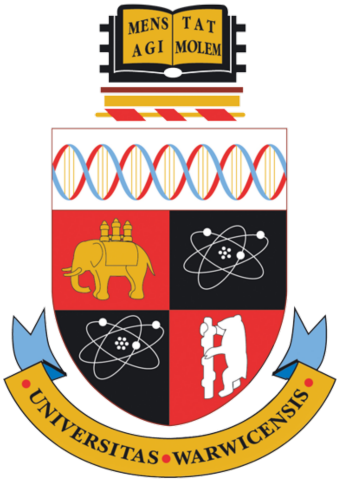
Correlation between prevalence of endemic infections in metapopulation networks

by

Sophie R. Meakin

Thesis submitted to the University of Warwick
for the degree of
Doctor of Philosophy



University of Warwick, Department of Mathematics

November 2019

Contents

Acknowledgements

I would like to thank everyone (and, believe me, there's a lot) who has given me a pep talk during the last four years. Many, many thanks go to my brilliant supervisor, Matt Keeling, for his support and encouragement, particularly of my foray into outbreak response. Thank you also to staff and students in MathSys and SBIDER (especially D1.02), and finally to my two very talented friends, April Waites and Claire Lancaster, for keeping me going and just being the best.

Declarations

The work presented here is my own, except where stated otherwise. This thesis has been composed by myself and has not been submitted for any other degree or professional qualification.

Chapter 2 and Appendix A have been published as

- S. Meakin and M. Keeling (2018). Correlations between stochastic epidemics in two interacting populations. *Epidemics.* doi:10.1016/j.epidem.2018.08.005.

Chapter 4 and Appendix B have been published as

- S. Meakin and M. Keeling (2019). Correlations between stochastic endemic infection in multiple interacting subpopulations. *Journal of Theoretical Biology.* doi: 10.1016/j.jtbi.2019.109991

Abstract

An ongoing challenge in metapopulation modelling of infectious diseases is how to infer the coupling, or level of interaction, between groups of individuals. The individual-level behaviour that determines the interactions between groups is highly complex, and good data on relevant interactions are not always readily available. Moreover, even with access to good data on relevant interactions, it is unclear how this should translate into a transmission parameter.

On the other hand, long-term data on disease incidence are often more widely available and can be used to estimate the correlation between infection prevalence in two interacting groups. In this thesis, we explore the relationship between the coupling and the correlation using two approximation methods from probability theory: moment closure approximations and diffusion approximations. We propose that this relationship can be used to infer the coupling from the observed infection incidence, even if the observations are limited in some way. We also use two methods to show how properties of the metapopulation network structure, such as degree and edge-density, affect the correlation.

They all say, Go on to graduate studies, and they give you a bit of money; so you do, and you think, Now I'm going to find out the real truth. But you don't find out, exactly, and things get pickier and pickier and more and more stale, and it all collapses in a welter of commas and shredded footnotes, and after a while it's like anything else: you've got stuck in it and you can't get out, and you wonder how you got there in the first place.

Margaret Atwood, The Edible Woman

Chapter 1

Background

1.1 Mathematical modelling of infectious diseases

Compartmental models are the most widely-used modelling framework for the study of infectious diseases, whereby individuals are classified into different compartments according to their disease status. Individuals then move through the various compartments at predefined rates such that the model describes the typical disease progression.

In the SIR model, individuals are in one of three states: susceptible (S), infected (and infectious; I), or recovered (R). Susceptible individuals meet other individuals at rate $k > 0$. We assume that these encounters are sufficiently close that if the other individual is infected, then transmission of infection occurs with probability τ and the susceptible individual immediately becomes infected and infectious to others. In standard epidemiological modelling notation, we let the transmission rate be $\beta = k\tau$. Infected individuals recover from infection at rate $\gamma > 0$, after which they become immune to further infection.

Additional compartments can be included depending on the disease being studied. Often there is a period of latent infection during which the individual is not yet infectious to others: this is included as an exposed (E) class, and creates the SEIR model. For diseases that confer no immunity (such as sexually-transmitted diseases), recovered individuals will return to the susceptible class and so creates the SIS model. Additional compartments may be added to account for asymptomatic infection (A), vaccination (V) or, in the case of vector-borne diseases, susceptible and infected vectors (S_{vec} and I_{vec}, respectively).

When the disease dynamics happen at a much faster rate than demographic events, then the effect of demographic events is negligible and can be ignored. However, for the study of endemic diseases then the introduction of new susceptible individuals is important, and so birth and death events should be included. In this case, we assume that individuals are born into the susceptible class at rate $\nu > 0$, and that individuals in all classes die at rate $\mu > 0$, independent of infection status. Note that both ν and μ are per capita rates.

Deterministic models

Compartmental models were originally studied as systems of ordinary differential equations (ODEs) describing the time evolution of the number or proportion of individuals in different compartments. The major contribution of Kermack and McKendrick (1927) was to write down the ODE system for the closed SIR model, given by:

$$\begin{aligned}
\frac{d}{dt}S(t) &= -\frac{\beta}{N}S(t)I(t) \\
\frac{d}{dt}I(t) &= \frac{\beta}{N}S(t)I(t) - \gamma I(t) \\
\frac{d}{dt}R(t) &= \gamma I(t),
\end{aligned}$$

where $S(t), I(t), R(t) \in [0, N]$ denote the number of susceptible, infected and recovered individuals, respectively, at time $t \geq 0$, and the total population size is N. Since demographic events are not included then the population size is constant, and so we can reduce the dimensionality of the system by setting $R(t) = N - S(t) - I(t)$.

An important ratio in the study of infectious diseases is the basic reproduction number R_0, defined as the average number of secondary cases generated by a single infected individual in an otherwise susceptible population. The value of R_0 is determined by the transmission rate β and the average length of the infectious period γ^{-1}. For the closed SIR model, $R_0 = \beta/\gamma$. At the start of an outbreak, the number of infected individuals is increasing if and only if $R_0 > 1$. This result follows from the fact that $dI(0)/dt$ is increasing if and only if $\beta S(0)/N > \gamma$ and $S(0) \approx N$. In general, then the number of infected individuals is increasing if and only if $S(t) > N/R_0$. The only equilibrium state (S^*, I^*, R^*) for the closed SIR model is the disease-free state, where $I^* = 0$.

When we add demographic events to the closed SIR model then the dynamics change

significantly. The ODE system for the SIR model with demographic events is given by

$$\frac{d}{dt}S(t) = \nu N - \frac{\beta}{N}S(t)I(t) - \mu S(t) \tag{1.1}$$

$$\frac{d}{dt}I(t) = \frac{\beta}{N}S(t)I(t) - \gamma I(t) - \mu I(t) \tag{1.2}$$

$$\frac{d}{dt}R(t) = \gamma I(t) - \mu R(t). \tag{1.3}$$

and so the basic reproduction number is $R_0 = \beta/(\gamma + \mu)$. It is clear that by including demographic events, the overall transition rate out of the infected class increases and so the average infectious period decreases. By introducing birth and death events we also allow for susceptible individuals to be replaced and so introduce a second equilibrium state $(S^*, I^*, R^*) = (N/R_0, N\mu(R_0 - 1)/\beta, N\gamma(R_0 - 1)/\beta)$. We refer to this as the endemic equilibrium. $R_0 > 1$ is a necessary condition for the existence of an endemic equilibrium state. Moreover, if $R_0 > 1$ then the endemic equilibrium state is stable; otherwise the disease-free equilibrium state is stable.

1.2 Stochastic models of infectious diseases

Although deterministic ODE models provide a simple and useful framework for the study of infectious diseases, the dynamics of an outbreak are clearly not deterministic. Stochastic effects are particularly important when the number of infected individuals is small: for example, whether or not an outbreak occurs as a result of a single new case depends on whether this individual infects other individuals before they recover.

Stochastic models of infectious diseases exhibit notable differences when compared to their deterministic counterparts. Under the same model parameters and starting conditions, repeated simulation of stochastic models gives rise to an ensemble of different realisations, whereas deterministic models predict only a single equilibrium solution with no deviation. As a result of stochastic fluctuations, it is possible for the disease to go locally extinct.

The most natural and flexible stochastic formulation of infectious disease dynamics is as a multivariable continuous-time Markov process. The state space of Markov processes is discrete, and so is appropriate to describe counts of individuals in different disease states. Under this framework, events (transmission, recovery, births, deaths, etc.) occur at the points of independent Poisson processes with rates defined by the current state of the system. The dimension of the Markov process is determined by number of tracked

Event	Transition	Rate
Infection	$s \to s-1, i \to i+1$	$\beta si/N$
Recovery	$i \to i-1, r \to r+1$	γi
Birth	$s \to s+1$	νN
Death (S)	$s \to s-1$	μs
Death (I)	$i \to i-1, i \to i-1$	μi
Death (R)	$r \to r-1, r \to r-1$	μr

Table 1.1. A summary of the transition rates of the two-dimensional Markov process SIR model $\{(S(t), I(t)) : t \geq 0\}$ from state (s, i) with transmission rate $\beta > 0$, recovery rate $\gamma > 0$, birth rate $\nu \geq 0$ and death rate $\mu \geq 0$.

compartments: for example, in the closed SIR model we track $S(t)$ and $I(t)$ only (and $R(t) = N - S(t) - I(t)$) and obtain a 2-dimensional Markov process $((S(t), I(t)), t \geq 0)$.

The continuous-time Markov process SIR model can be described as follows. Susceptible individuals meet infected individuals at points of a Poisson process with rate $\beta I/N$, and infected individuals remain infected for an exponentially distributed time with mean γ^{-1}. If demographic events are included, then individuals are born into the susceptible class at rate ν and are alive for an exponentially distributed time with mean μ^{-1}. These transition rates are summarised in Table 1.1. The closed SIR model (without demography) is a 2-dimensional continuous-time Markov process $((S(t), I(t)), t \geq 0)$; the SIR model with demographic events is a 3-dimensional continuous time Markov process $((S(t), I(t), R(t)), t \geq 0)$.

The Markov process is obviously not the only method to describe a stochastic infectious disease processes. A thorough overview of the full breadth of stochastic infectious disease modelling is given by Bailey (1975), Anderson and May (1992), Andersson and Britton (2000), Diekmann and Heesterbeek (2000) and Keeling and Rohani (2008); a very brief summary is included here. In the early stages of an epidemic the Markov process can be approximated by a linear birth-death process (Grimmet and Stirzaker, 2001; Andersson and Britton, 2000), from which we can calculate the probability of a major outbreak; note that this result only holds for $I(t) = o(\sqrt{N})$. When $R_0 > 1$ we can ask what fraction z of the population will be affected by a major outbreak: the Sellke construction can be used to show that this is the solution to $1 - z = e^{-R_0 z}$ (Sellke, 1983). In the large population limit, the Markov process can be approximated by a Gaussian diffusion process around the deterministic endemic equilibrium (Kurtz, 1970, 1971; Barbour, 1972, 1974); the time evolution of the probability density function can

be described by a Fokker-Planck equation (Nasell, 1999) and, around the deterministic endemic equilibrium, an Ornstein-Uhlenbeck process. The deterministic ODE system is therefore a large-population approximation of the stochastic process. In this thesis we consider only continuous-time Markov processes (as described above) and the diffusion approximation of this process (see Section 1.2.3).

1.2.1 Kolmogorov forward equations for Markov processes

The Kolmogorov forward equations describes continuous time-evolution of the probability density function of a Markov process. These, along with the Kolmogorov backward equations, were first described by Kolmogoroff (1931); an analogous concept for diffusion processes (continuous, rather than discrete, state space) is known within physics as the Fokker-Planck equation (see Section 1.2.3). Here we recall the definition of the Kolmogorov forward equations for a multivariable stochastic process.

Let $(\boldsymbol{X}_t, t \geq 0)$ be a K-dimensional stochastic process with state space $\mathbb{N}^K$. Let $p_t(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X}_t = \boldsymbol{x})$ denote the probability that $\boldsymbol{X} = \boldsymbol{x}$ at time t, and let $W(\boldsymbol{x}, \boldsymbol{r}), \boldsymbol{r} \in \mathbb{Z}^K$ be the rate at which the process 'jumps' from state $\boldsymbol{x}$ to state $\boldsymbol{x} + \boldsymbol{r}$. The Kolmogorov forward equation for this process can be written as

$$\frac{dp_t(\boldsymbol{x})}{dt} = \sum_{\boldsymbol{r}} [W(\boldsymbol{x} - \boldsymbol{r}, \boldsymbol{r}) p_t(\boldsymbol{x} - \boldsymbol{r}) - W(\boldsymbol{x}, \boldsymbol{r}) p_t(\boldsymbol{x})]. \tag{1.4}$$

For the closed SIR model the Kolmogorov forward equations are given by

$$\frac{dp_t(s,i)}{dt} = \frac{\beta}{N}(s+1)(i-1)p_t(s+1,i-1) + \gamma(i+1)p_t(s,i+1) - \left(\frac{\beta}{N}si + \gamma i\right) p_t(s,i), \tag{1.5}$$

where $p_t(s,i) = \mathbb{P}(S(t) = s, I(t) = i)$ denotes the probability that there are s susceptible individuals and i infectious individuals in the population at time t.

The Kolmogorov forward equation formulation is extremely useful for describing the complete nature of a given stochastic system. Models of infectious disease dynamics are particularly amenable to this approach since the space of possible transitions is very small (e.g. to the current state ± 1). For small population sizes and simple epidemic dynamics (e.g. SIS or SIR) the Kolmogorov forward equations can be solved directly (Jacquez and Simon, 1993; Keeling and Ross, 2008). However, this approach quickly becomes infeasible for larger population sizes or more biologically-realistic models. For a general process with c disease compartments and a population of size N, the number

of equations grows like $N^c/c!$ (Keeling and Ross, 2008). For example, for the closed SIR model, there are $(N+1)(N+2)/2$ equations; when $N = 100$, we have 5151 equations to solve. For the closed SEIR model (that is, where $R(t) = N - S(t) - E(t) - I(t)$), when $N = 100$, we have $176,851$ equations.

Two alternative approaches can be taken. The first is to simulate multiple realisations of the stochastic model using some algorithm. The second is to study the expected behaviour of the model by making some analytic approximation: here we consider moment-closure approximations and diffusion approximations. In the former, a closed system of ODEs is constructed that describes the expected behaviour of the first- and second-order moments of the stochastic process. In the latter, the Markov process is approximated by a Gaussian diffusion process. In both approaches some loss of precision is made for analytical tractability.

1.2.2 Deterministic behaviour of moments

As discussed above, exact analysis of Markovian infectious disease models is often mathematically intractable. However, instead we can consider the deterministic behaviour of the first- and second-order moments of the process. Beginning with the Kolmogorov forward equation, we can write down an ODE for each of the time evolution of first- and second-order moments. In summary, the ODE for $\mathbb{E}[X_i^m X_j^n]$ can be written as

$$\frac{d\mathbb{E}[X_i^m X_j^n]}{dt} = \mathbb{E}\left[\sum_{events} \text{rate of event} \times \text{change in } X_i^m X_j^n \text{ due to event}\right],$$

and thus can be easily calculated from a list or table of transition rates and changes.

In this section we outline the derivation of this equation: we begin with moments of the form $\mathbb{E}[X^n]$ for single variable processes; this can then be extended to moments of the form $\mathbb{E}[X_i^m X_j^n]$ for multivariable stochastic processes. In this way we can derive ODEs for all first- and second-order moments (and so also for the variances and covariances) for any Markovian model of infectious disease dynamics.

Let $(X(t), t \geq 0)$ be a continuous-time Markov process with state space $\mathbb{N} = \{0, 1, 2, \ldots\}$. Let $p_t(x) = \mathbb{P}(X(t) = x)$, and let $W(x, r), r \in \mathbb{Z}$ be the rate at which the process 'jumps' from state x to state $x + r$. We show that the ODE for the time evolution of moments

of the form $\mathbb{E}[X^n], n \in \mathbb{N}$, is given by

$$\frac{d\mathbb{E}[X^n]}{dt} = \mathbb{E}\left[\sum_r [(X+r)^n - X^n] W(X,r)\right]. \tag{1.6}$$

We multiply both sides of the Kolmogorov forward equation by x^n and sum over $x \geq 0$. On the left hand side we have

$$\sum_{x=0}^{\infty} x^n \frac{dp_t(x)}{dt} = \frac{d}{dt}\sum_{x=0}^{\infty} x^n p_t(x) = \frac{d\mathbb{E}[X^n]}{dt},$$

and on the right hand side we have

$$\begin{aligned}
&\sum_{x=0}^{\infty} x^n \sum_r [W(x-r,r)p_t(x-r) - W(x,r)p_t(x)] \\
&= \sum_{x=0}^{\infty}\sum_r x^n W(x-r,r)p_t(x-r) - \sum_{x=0}^{\infty}\sum_r x^n W(x,r)p_t(x) \\
&= \sum_{x=-r}^{\infty}\sum_r (x+r)^n W(x,r)p_t(x) - \sum_{x=0}^{\infty}\sum_r x^n W(x,r)p_t(x) \\
&= \sum_{x=0}^{\infty} p_t(x) \sum_r [(x+r)^n - x^n] W(x,r) \\
&= \mathbb{E}\left[\sum_r [(X+r)^n - X^n] W(X,r)\right].
\end{aligned}$$

It is straightforward to extend this result to a multivariable Markov processes. Let $(\boldsymbol{X}(t), t \geq 0)$ be an K-dimensional continuous-time Markov process with state space $\mathbb{N}^K$, and let $W(\boldsymbol{x}, \boldsymbol{r})$ be the transition rate from state $\boldsymbol{x}$ to state $\boldsymbol{x} + \boldsymbol{r}$. We show that the ODE for the time evolution of moments of the form $\mathbb{E}[X_i^m X_j^n], m, n \in \mathbb{N}$ is given by

$$\frac{d\mathbb{E}[X_i^m X_j^n]}{dt} = \mathbb{E}\left[\sum_{\boldsymbol{r}} [(X_i + r_i)^m (X_j + r_j)^n - X_i^m X_j^n] W(\boldsymbol{X}, \boldsymbol{r})\right].$$

For infectious disease processes, the probability mass function $p_t(\boldsymbol{x})$ is defined only for non-negative values of $\boldsymbol{x}$, that is where $x_k \geq 0, \forall k$; for simplicity we write this as $\boldsymbol{x} \geq \boldsymbol{0}$. Similar to the single variable case, we multiply both sides of the Kolmogorov

forward equation by $x_i^m x_j^n$ and sum over $\boldsymbol{x}$ on both sides. On the left hand side we have

$$\sum_{\boldsymbol{x}\geq\boldsymbol{0}} x_i^m x_j^n \frac{dp_t(\boldsymbol{x})}{dt} = \frac{d}{dt}\sum_{\boldsymbol{x}\geq\boldsymbol{0}} x_i^m x_j^n p_t(\boldsymbol{x}) = \frac{d\mathbb{E}[X_i X_j]}{dt},$$

and on the right-hand side we have

$$\begin{aligned}
&\sum_{\boldsymbol{x}\geq\boldsymbol{0}} x_i^m x_j^n \sum_{\boldsymbol{r}} [W(\boldsymbol{x}-\boldsymbol{r},\boldsymbol{r})p_t(\boldsymbol{x}-\boldsymbol{r}) - W(\boldsymbol{x},\boldsymbol{r})p_t(\boldsymbol{x})] \\
&= \sum_{\boldsymbol{x}\geq\boldsymbol{0}}\sum_{\boldsymbol{r}} x_i^m x_j^n W(\boldsymbol{x}-\boldsymbol{r},\boldsymbol{r})p_t(\boldsymbol{x}-\boldsymbol{r}) - \sum_{\boldsymbol{x}\geq\boldsymbol{0}}\sum_{\boldsymbol{r}} x_i^m x_j^n W(\boldsymbol{x},\boldsymbol{r})p_t(\boldsymbol{x}) \\
&= \sum_{\substack{\boldsymbol{x}: \\ x_i\geq -r_i, x_j\geq -r_j}} \sum_{\boldsymbol{r}} (x_i+r_i)^m (x_j+r_j)^n W(\boldsymbol{x},\boldsymbol{r})p_t(\boldsymbol{x}) - \sum_{\boldsymbol{x}\geq\boldsymbol{0}}\sum_{\boldsymbol{r}} x_i^m x_j^n W(\boldsymbol{x},\boldsymbol{r})p_t(\boldsymbol{x}) \\
&= \sum_{\boldsymbol{x}\geq\boldsymbol{0}}\sum_{\boldsymbol{r}} (x_i+r_i)^m (x_j+r_j)^n W(\boldsymbol{x},\boldsymbol{r})p_t(\boldsymbol{x}) - \sum_{\boldsymbol{x}\geq\boldsymbol{0}}\sum_{\boldsymbol{r}} x_i^m x_j^n W(\boldsymbol{x},\boldsymbol{r})p_t(\boldsymbol{x}) \\
&= \sum_{\boldsymbol{x}\geq\boldsymbol{0}} p_t(\boldsymbol{x}) \sum_{\boldsymbol{r}} [(x_i+r_i)^m (x_j+r_j)^n - x_i^m x_j^n] W(\boldsymbol{x},\boldsymbol{r}) \\
&= \mathbb{E}\left[\sum_{\boldsymbol{r}} [(X_i+r_i)^m (X_j+r_j)^n - X_i^m X_j^n] W(\boldsymbol{X},\boldsymbol{r})\right].
\end{aligned}$$

We can use this method to derive ODEs for any first- and second-order moments of a stochastic infectious disease mode. For the closed SIR model, there are two first-order moments ($\mathbb{E}[S]$ and $\mathbb{E}[I]$) and three second-order moments ($\mathbb{E}[S^2], \mathbb{E}[I^2]$ and $\mathbb{E}[SI]$). The ODEs for the first order moments are

$$\begin{aligned}
\frac{d\mathbb{E}[S]}{dt} &= \mathbb{E}\left[(-1)\frac{\beta SI}{N}\right] = -\frac{\beta\mathbb{E}[SI]}{N} \\
\frac{d\mathbb{E}[I]}{dt} &= \mathbb{E}\left[(+1)\frac{\beta SI}{N} + (-1)\gamma I\right] = \frac{\beta\mathbb{E}[SI]}{N} - \gamma\mathbb{E}[I],
\end{aligned}$$

and the ODEs for the second-order moments are

$$\begin{aligned}
\frac{d\mathbb{E}[S^2]}{dt} &= \mathbb{E}\left[(-2S+1)\frac{\beta SI}{N}\right] = -\frac{\beta}{N}\left[2\mathbb{E}[S^2 I] - \mathbb{E}[SI]\right] \\
\frac{d\mathbb{E}[I^2]}{dt} &= \mathbb{E}\left[(2I+1)\frac{\beta SI}{N} + (-2I+1)\gamma I\right] = \frac{\beta}{N}\left[2\mathbb{E}[SI^2] + \mathbb{E}[SI]\right] - \gamma[2\mathbb{E}[I^2] - \mathbb{E}[I]] \\
\frac{d\mathbb{E}[SI]}{dt} &= \mathbb{E}\left[(S-I-1)\frac{\beta SI}{N} + (-S)\gamma I\right] = \frac{\beta}{N}\left[\mathbb{E}[S^2 I] - \mathbb{E}[SI^2] - \mathbb{E}[SI]\right] - \gamma\mathbb{E}[SI].
\end{aligned}$$

Moment closure approximations

Due to the non-linearity of the infection term in infectious disease models, the ODE for an nth-order moment will depend on one or more $(n+1)$th-order moments. For example, in the closed SIR model, the ODE for $\mathbb{E}[S]$ contains a $\mathbb{E}[SI]$ term, and the ODE for $\mathbb{E}[SI]$ contains both $\mathbb{E}[S^2I]$ and $\mathbb{E}[SI^2]$ terms, and so on. Therefore, to fully define the system of ODEs we would have to write down an infinite set of equations. The usual approach to circumvent this problem is to assume that the distribution of states follow some known distribution, then use the relationship between moments to truncate the set of ODEs at some order. This method is called a moment closure approximation, since it is used to close the equations for the moments.

The most commonly used moment closure approximation, and the one used throughout this thesis, assumes that the distribution of states follows a multivariate normal distribution (Whittle, 1957; Isham, 1991, 1993; Keeling and Rohani, 2002; Lloyd, 2004). This approximation holds in the large-population limit (Kurtz, 1970, 1971; Nasell, 1999; Lloyd, 2004), and fails when there are large negative covariances, frequent global extinctions or when the distribution of states is bimodal (Nasell, 1999; Keeling, 2000a,b; Lloyd, 2004; Krishnarajah et al., 2005). Under the multivariate normal assumption, third-order central moments are equal to zero, and so third-order moments can simply be written in terms of first and second order moments:

$$\begin{aligned} 0 &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])(Z-\mathbb{E}[Z])] \\ \iff 0 &= \mathbb{E}[XYZ]-\mathbb{E}[X]\mathbb{E}[YZ]-\mathbb{E}[Y]\mathbb{E}[XZ]-\mathbb{E}[Z]\mathbb{E}[XY]+2\mathbb{E}[X]\mathbb{E}[Y]\mathbb{E}[Z] \\ \iff \mathbb{E}[XYZ] &= \mathbb{E}[X]cov(Y,Z)+\mathbb{E}[Y]cov(X,Z)+\mathbb{E}[Z]cov(X,Y)+\mathbb{E}[X]\mathbb{E}[Y]\mathbb{E}[Z]. \end{aligned}$$

Alternative approximations have been proposed based on different distributional assumptions. The multiplicative moment closure approximation proposed by Keeling (2000a,b) assumes a multivariate log-normal distribution; this is only defined for positive values and so avoids the problem of negative average densities that can arise with the multivariate normal moment closure approximation. Nasell (2003a,b) considers the effects of approximating the stationary distribution by distributions that are themselves approximations for the Normal distribution: a Poisson distribution, a log-normal distribution and a binomial distribution. Finally, Krishnarajah et al. (2005) assumes a beta-binomial distribution.

In Chapter 2 and 4 we use the multivariate normal moment closure approximation to

write down a closed system of ODEs describing the dynamics of an SIR-type model in a metapopulation network. The aim of this approach is to study the correlation between infection prevalence in different subpopulations.

1.2.3 Diffusion approximation

In the large population limit, the Markov process can be approximated by a Gaussian diffusion process around the deterministic endemic equilibrium (Kurtz, 1970, 1971; Barbour, 1972, 1974). The Fokker-Planck equation describes the time evolution of the probability density function of such processes (Nasell, 1999), and is the continuous-state analogue to the Kolmogorov forward equation for discrete-state Markov processes. We recall the definition of the Fokker-Planck equation for a multivariable diffusion process and some basic results of the expected behaviour of the first- and second-order moments.

Let $(\boldsymbol{X}(t), t \geq 0)$ be a K-dimensional continuous-time Markov process with state space $\mathbb{R}^K$ and let $W(\boldsymbol{x}, \boldsymbol{r})$ be the transition rate from state $\boldsymbol{x}$ to state $\boldsymbol{x} + \boldsymbol{r}$. The multidimensional Fokker-Planck equation for this process is

$$\frac{\partial P}{\partial t} = -\sum_{i=1}^{K} \frac{\partial}{\partial x_i}(A_i P) + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\partial^2}{\partial x_i \partial x_j}(B_{ij} P), \tag{1.7}$$

where

$$A_i(\boldsymbol{x}) = \sum_{\boldsymbol{r}} r_i W(\boldsymbol{x}, \boldsymbol{r}) \tag{1.8}$$

$$B_{ij}(\boldsymbol{x}) = \sum_{\boldsymbol{r}} r_i r_j W(\boldsymbol{x}, \boldsymbol{r}). \tag{1.9}$$

To derive Equation (1.7) we begin with the Kolmogorov forward equation:

$$\frac{dp_t(\boldsymbol{n})}{dt} = \sum_{\boldsymbol{r}} [W(\boldsymbol{n} - \boldsymbol{r}, \boldsymbol{r}) p_t(\boldsymbol{n} - \boldsymbol{r}) - W(\boldsymbol{n}, \boldsymbol{r}) p_t(\boldsymbol{n})],$$

where $\boldsymbol{n} \in \mathbb{N}^K$. The right-hand side of this equation is of the form $f(\boldsymbol{n}-\boldsymbol{r})-f(\boldsymbol{n})$, where $f(\boldsymbol{n}) \coloneqq W(\boldsymbol{n}, \boldsymbol{r}) P(\boldsymbol{n}, t)$; we consider the second-order Taylor expansion of $f(\boldsymbol{x}-\boldsymbol{r})-f(\boldsymbol{x})$ around $\boldsymbol{x}$ (i.e. we assume that $|\boldsymbol{r}| \ll |\boldsymbol{x}|$, which holds in the large-population limit) and

so have

$$\begin{aligned}
\frac{\partial P(\boldsymbol{x},t)}{\partial t} &= \sum_{\boldsymbol{r}}[W(\boldsymbol{x}-\boldsymbol{r},\boldsymbol{r})p_t(\boldsymbol{x}-\boldsymbol{r}) - W(\boldsymbol{x},\boldsymbol{r})p_t(\boldsymbol{x})] \\
&= \sum_{\boldsymbol{r}}[-\sum_{i=1}^{K} r_i \frac{\partial}{\partial x_i}(W(\boldsymbol{x},\boldsymbol{r})P(\boldsymbol{x},t)) + \frac{1}{2}\sum_{i=1}^{K} r_i r_j \frac{\partial^2}{\partial x_i \partial x_j}(W(\boldsymbol{x},\boldsymbol{r})P(\boldsymbol{x},t))] \\
&= -\sum_{i=1}^{K} \frac{\partial}{\partial x_i}\left(\left(\sum_{\boldsymbol{r}} r_i W(\boldsymbol{x},\boldsymbol{r})\right)P(\boldsymbol{x},t)\right) + \frac{1}{2}\sum_{i=1}^{K} \frac{\partial^2}{\partial x_i \partial x_j}\left(\left(\sum_{\boldsymbol{r}} r_i r_j W(\boldsymbol{x},\boldsymbol{r})\right)P(\boldsymbol{x},t)\right) \\
&= -\sum_{i=1}^{K} \frac{\partial}{\partial x_i}(A_i P) + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\partial^2}{\partial x_i \partial x_j}(B_{ij}P).
\end{aligned}$$

The behaviour of the mean and covariances for this process can be described in terms of $\boldsymbol{A}$ and $\boldsymbol{B}$. The time evolution of the mean is given by

$$\frac{d}{dt}\mathbb{E}[X_m] = A_m,$$

that is, the drift of the mean is determined entirely by the matrix $\boldsymbol{A}$. The time evolution of $\mathbb{E}[X_m X_n]$ is given by

$$\frac{d}{dt}\mathbb{E}[X_m X_n] = \mathbb{E}[A_m X_n] + \mathbb{E}[A_n X_m] + \mathbb{E}[B_{mn}],$$

and so, if $\boldsymbol{C}$ is the covariance matrix with mn-th element $C_{mn} = cov(X_m, X_n)$, the time evolution of the covariance C_{mn} is given by

$$\begin{aligned}
\frac{d}{dt}C_{mn} &= \frac{d}{dt}\mathbb{E}[X_m X_n] - X_m \frac{d}{dt}\mathbb{E}[X_n] - X_n \frac{d}{dt}\mathbb{E}[X_m] \\
&= \mathbb{E}[A_m X_n] - X_m A_n + \mathbb{E}[A_n X_m] - X_n A_m + \mathbb{E}[B_{mn}] \\
&= cov(A_m, X_n) + cov(A_n, X_m) + \mathbb{E}[B_{mn}]. \qquad (1.10)
\end{aligned}$$

Dynamics at endemic equilibrium

The expected behaviour of the process at endemic equilibrium is found by solving $A_m = 0$ for $m = 1, \ldots, K$. What we are really interested in is the behaviour of the covariance $C_{mn} = cov(X_m, X_n)$ at endemic equilibrium. We approximate the behaviour of the stochastic process near endemic equilibrium so that A_m is linear in $\boldsymbol{x}$ and B_{mn} is constant

in $\boldsymbol{x}$:

$$A_m(\boldsymbol{x}) = \lambda_m + \sum_k a_{mk} x_k \tag{1.11}$$

$$B_{mn}(\boldsymbol{x}) = b_{mn}, \tag{1.12}$$

where the matrix $\boldsymbol{a}$ is simply the Jacobian of $\boldsymbol{A}$ at endemic equilibrium:

$$a_{mk} = \frac{\partial A_m}{\partial x_k}(\boldsymbol{x}^*).$$

We substitute these approximations for $\boldsymbol{A}$ and $\boldsymbol{B}$ into Equation (1.10) and get

$$\begin{aligned}
\frac{d}{dt}C_{mn} &= cov(\lambda_m + \sum_k a_{mk}X_k, X_n) + cov(\lambda_n + \sum_k a_{nk}X_k, X_m) + b_{mn} \\
&= \sum_k a_{mk} cov(X_k, X_n) + \sum_k a_{nk} cov(X_k, X_m) + b_{mn} \\
&= \sum_k a_{mk} C_{kn} + \sum_k a_{nk} C_{km} + b_{mn},
\end{aligned}$$

or, equivalently, in matrix form

$$\frac{d\boldsymbol{C}}{dt} = \boldsymbol{aC} + \boldsymbol{C}\boldsymbol{a}^T + \boldsymbol{b}.$$

At endemic equilibrium[1] $d\boldsymbol{C}/dt = 0$ and so the covariance matrix, $\boldsymbol{C}$, is the solution to the Lyapunov equation:

$$\sum_k a_{mk} C_{kn} + \sum_k a_{nk} C_{km} = -b_{mn} \tag{1.13}$$

$$\iff \boldsymbol{aC} + \boldsymbol{C}\boldsymbol{a}^T = -\boldsymbol{b}. \tag{1.14}$$

1.2.4 Simulation of Markov processes

Some stochastic models of infectious diseases may be too complex to study analytically. Given technological advances in computational power, we can instead simulate the stochastic process and analyse the numerical results in lieu of studying the process analytically. An advantage of this is that we are able to evaluate the accuracy of analytic

[1] We acknowledge that this is not strictly endemic equilibrium, but endemic stationarity. However, in infectious disease modelling, these two terms are used interchangeably and we will follow this convention throughout the thesis.

approximation methods: for example, by comparing simulation results to the analytic results derived by using the moment closure approximation from Section 1.2.2 or the diffusion approximation from Section 1.2.3, we can easily see the effect of making these approximations.

However, for this method to be viable, we need to be able to simulate the stochastic process effectively. It is unclear whether a single realisation is representative of the average behaviour of the system, or whether we have actually observed an unlikely trajectory by chance. To be sure that we are truly observing the expected behaviour of the Markov process, we need to be able to generate many realisations. We describe two stochastic simulation algorithms used widely within infectious disease modelling, and used in the main chapters of this thesis.

Gillespie algorithm

Gillespie's stochastic simulation algorithm, simply known as the Gillespie algorithm, (Gillespie, 1976, 1977) is widely used in infectious disease modelling to generate individual realisations of stochastic systems.

After defining the initial state of the stochastic process, the algorithm repeats a Monte Carlo step until the defined maximum time is exceeded (see Algorithm 1 for pseudocode). The Monte Carlo step is split into two parts. First, the algorithm determines the time to the next event which is exponentially distributed with rate equal to the sum of the rates of all possible events. Second, the algorithm determines which event occurs next: the probability that an event occurs is proportional to the rate of that event. A single realisation of the Gillespie algorithm represents a sample from the probability mass function that is the solution to the Kolmogorov forward equation. However, the algorithm is computationally expensive, and for large populations or systems with many possible events, the Gillespie algorithm becomes less efficient. This is because as the number of particles and/or transitions increases, then the total rate $R(\boldsymbol{x}(t))$ increases and the expected time to the next event, $1/R(\boldsymbol{x}(t))$, becomes very small. Many modifications of the original Gillespie algorithm exist to address this problem (Gillespie, 2001; Gillespie and Petzold, 2003; Rathinam et al., 2003; Cao et al., 2006; Gillespie, 2007).

Algorithm 1: Gillespie's stochastic simulation algorithm (Gillespie, 1976, 1977)

Initialization: specify T_{max}, set $t = 0$, define $\boldsymbol{x}(0) = \boldsymbol{x}_0$;
while $t < T_{max}$ **do**
  Compute rates of events $R_i(\boldsymbol{x}(t)), i = 1 \ldots, N$;
  $R(\boldsymbol{x}(t)) = \sum_i R_i(\boldsymbol{x}(t))$;
  $\delta t \sim \text{Exp}(R(\boldsymbol{x}(t))$;
  $r \sim \text{Uniform}[0, 1]$;
  $a = rR(\boldsymbol{x}(t))$;
  if $\sum_{i=1}^{j-1} R_i(\boldsymbol{x}(t)) < a < \sum_{i=1}^{j} R_i(\boldsymbol{x}(t))$ **then**
    Event j occurs;
  Update $\boldsymbol{x}(t)$;
  $t = t + \delta t$;

τ-leaping algorithm

The τ-leaping algorithm is an approximation of the Gillespie algorithm that can improve the simulation speed in exchange for some loss in precision. Gillespie (2001) describes this as 'leaping' along the system's history axis, rather than 'stepping' as in the direct Gillespie algorithm, hence the name of the algorithm.

The approach of the algorithm is as follows: instead of observing the system at the time of every individual reaction, we observe the system at time intervals of length τ and estimate how many of each reaction occured in that time interval (see Algorithm 2 for pseudocode). As multiple events can occur at once, it is neccessary to check that the new state of the system is physically allowed; for models of infectious disease dynamics, then we should also ensure $x_i(t) \leq N$, in addition to non-negativity. The time step τ is chosen such that the change in state during $(t, t + \tau)$ is small and so the event rates are essentially constant and equal to $R_i(\boldsymbol{x}(t)), i = 1, \ldots, N$. A substantial body of literature has developed concerning the choice of τ. In the simplest case τ is fixed; alternatively τ is determined as a function of the event rates (Gillespie, 2001; Gillespie and Petzold, 2003; Rathinam et al., 2003; Cao et al., 2006; Gillespie, 2007).

1.3 Heterogeneity in infectious diseases models

One of the main assumptions of many traditional infectious disease models is homogeneity: individuals are identical and homogeneously mixing, that is, each individual in the

Algorithm 2: τ-leaping algorithm (Gillespie, 2001)

Initialization: specify T_{max}, set $t = 0$, define $\boldsymbol{x}(0) = \boldsymbol{x}_0$;
while $t < T_{max}$ **do**
  Compute rate of event i: $R_i(\boldsymbol{x}(t))$;
  Choose time step τ;
  Compute occurence of event i in $[t, t+\tau)$: $k_i \sim \text{Poisson}(R_i(t)\tau)$;
  Update $\boldsymbol{x}(t)$;
  Ensure non-negativity of $\boldsymbol{x}(t)$: $x_i(t) = \max(0, x_i(t))$;
  $t = t + \tau$;

population is equally likely to have contact with all other individuals. Both assumptions are made for mathematical tractability, but it is intuitive that they are a major over simplification. Individuals may vary in their susceptibility to a disease, and this has been shown to have a marked effect on the dynamics of infectious diseases (Rodrigues et al., 2009; Capała and Dybiec, 2017). In this thesis, and the rest of this section, we discuss the limitations of the homogeneous mixing assumption.

Many studies have confirmed the intuition that homogeneous mixing is unrealistic, showing that social contact patterns are highly heterogeneous (Mossong et al., 2008; González et al., 2008; Horby et al., 2011; Danon et al., 2013; Read et al., 2014; Stopczynski et al., 2014; Wesolowski et al., 2015; Kiti et al., 2016; Klepac et al., 2018). Social contact networks often exhibit assortative mixing, such that contact is more frequently made within, rather than outside of, demographic groups. This includes (but is by no means limited to) assortative mixing by: age (Mossong et al., 2008; Danon et al., 2013; Read et al., 2014; Kiti et al., 2016), gender (Cauchemez et al., 2011; Scott et al., 2012; Stehlé et al., 2013), sexual identity (Schneider et al., 2013), and drug use (Schneider et al., 2013). Moreover, studies consistently show that the number and duration of contacts that individuals make is highly variable, and is also influenced by socio-demographic factors such as age, gender and occupation. Individuals usually make repeated contact with the same people, and these contacts are more likely to be physical and last for a longer time than compared to new contacts (Mossong et al., 2008; Horby et al., 2011). Finally, the majority of contacts are usually made in a small number of locations close to home (González et al., 2008; Klepac et al., 2018).

Heterogeneity has marked influences on many properties of population-level infectious disease dynamics. Individuals with large numbers of contacts can result in so called super-spreading events where a single infected individual generates a large num-

ber of secondary cases. Superspreading events play a key role in sustaining onward transmission during outbreaks of Ebola virus disease (Lau et al., 2017), severe acute respiratory syndrome (SARS) (Lloyd-Smith et al., 2005), Middle Eastern respiratory syndrome (MERS) (Kucharski and Althaus, 2015) and measles (De Serres et al., 2013), amongst others (Lloyd-Smith et al., 2005), as well as triggering outbreaks when following an importation event (De Serres et al., 2013). Variance in number of contacts also plays a role in early transmission dynamics (Anderson and May, 1992). As a result of structured social contact, cases are often clustered spatially or within high-risk demographics (Wu et al., 2004; Raymond and McFarland, 2009; Cauchemez et al., 2011; Schneider et al., 2013; Gog et al., 2014). Heterogeneity generally acts to increase persistence of diseases within stochastic populations (Lloyd and May, 1996; Keeling, 2000a,b; Hagenaars et al., 2004; Lloyd and Jansen, 2004), and can have both a positive and negative effect on the control of a disease: on the one hand, heterogeneity allows for targeted interventions (Christley et al., 2005; Keeling and White, 2010; Wallinga et al., 2010); on the other hand, it can lead to systematic non-adherence for treatment or vaccination (Dyson et al., 2017).

The main challenge of incorporating heterogeneity into infectious disease models is a balancing act between realism and mathematical amenability. At one extreme lie homogeneous-mixing models; as discussed earlier, these models benefit from being mathematically tractable, but also fail to capture some essential behaviour. At the other extreme are individual-based models, where each individual in the population is modelled separately and has unique individual-level attributes that determine their behaviour. The parametrisation and analysis of such models is challenging and computationally intensive, or impossible if there is insufficient data.

Between these extremes are network models and metapopulation models, which present some sort of compromise of the two approaches. Both models use networks to capture different levels of heterogeneity in contact patterns. In network models, nodes represent individuals and edges represent contact between individuals. A review of the extensive literature of network models is given by Danon et al. (2011). In this thesis, however, we are concerned with metapopulation models, which we discuss in depth in the following section.

1.3.1 Metapopulation models

Metapopulation theory has its roots in ecology, where it is used to consider the processes of regional extinction and recolonisation in spatially separated subpopulations connected by migration (Levins, 1969; Hanski and Gilpin, 1991; Hanski, 1998; Hanski and Simberloff, 1997). This framework has proved to be equally useful to capture multiple forms of heterogeneity in infectious disease modelling. In such models, the population is divided into multiple interacting, or 'coupled', subpopulations, where within-population interactions typically occur at a much higher rate than between-population interactions. Metapopulation models are most often considered as a form of spatial model, where subpopulations represent spatially separated groups (such as households, towns, cities, regions, countries); however, this framework can equally well apply to age or risk structured mixing, sexual mixing, or to multiple host species.

How interaction between subpopulations is incorporated into models varies, although broadly models are either mechanistic or phenomenological. Mechanistic metapopulation models explicitly describe movement (either commuting or migration) between subpopulations (Sattenspiel and Dietz, 1995; Keeling and Rohani, 2002; Jesse et al., 2008; Balcan et al., 2009, 2010; Belik et al., 2011; Gog et al., 2014). Technological advances have generated increasingly large spatiotemporal data on human behaviour, mobility and demography; in turn, this has allowed the development of increasingly realistic data-driven models. For example, the Global Epidemic and Mobility (GLEaM) model integrates airline, commuting and demographic data to simulate the spread of epidemics at the global scale (Balcan et al., 2009, 2010). Phenomenological models, on the other hand, simply express the force of infection as a function of infection prevalence in other subpopulations (Keeling and Rohani, 2002; Hagenaars et al., 2004; Kraemer et al., 2017; Hilton and Keeling, 2019). A strength of this simpler approach is that such models are more mathematically tractable, and so can be analysed both analytically and numerically.

Challenges in metapopulation infectious disease modelling

Quantifying between-population interactions is one of the key challenges of metapopulation infectious disease modelling (Ball et al., 2014). The interaction between subpopulations is often represented as a matrix of transmission rates within and between subpopulations, which has clear links to the number of cases generated by each group and hence to the basic reproductive ratio, R_0, through the dominant eigenvalue (Diekmann

et al., 1990; Heesterbeek, 2002). When dealing with P subpopulations, this transmission matrix has P^2 terms, which creates unidentifiability problems when attempting to estimate parameters from endemic equilibria, as we only have P pieces of information (Grenfell and Anderson, 1985).

As already discussed, the individual-level behaviour that determines the interactions between groups is highly complex and is dependent on socio-demographic factors (Mossong et al., 2008; González et al., 2008; Horby et al., 2011; Danon et al., 2013; Read et al., 2014; Stopczynski et al., 2014; Wesolowski et al., 2015; Kiti et al., 2016; Klepac et al., 2018). Even with access to good data on relevant interactions, it is unclear how this should translate into a transmission parameter. Moreover, good data on relevant interactions between subpopulations are rare, although technological developments mean that it is increasingly feasible to collect relevant data in novel ways. Mobile phone apps have been used to capture individual movements and contacts through specifically designed apps (Stopczynski et al., 2014; Klepac et al., 2018), or by using telephone calls as a proxy for human mobility (Wesolowski et al., 2015). Wearable sensors have also been used to gather data on interactions in small populations (Cattuto et al., 2010; Stehlé et al., 2013; Kiti et al., 2016).

Spatial metapopulation models rely heavily upon theoretical models of human mobility, which characterise the distribution of contacts between subpopulations based on the subpopulation sizes and the distances between them (Hanski, 1998). Such models are fit with appropriate interaction or mobility data, such as airline traffic data (Balcan et al., 2009, 2010), commuter mobility data (Viboud et al., 2006; Balcan et al., 2009, 2010), or mobile phone data, used as a proxy for human mobility (Tizzoni et al., 2014; Wesolowski et al., 2015; Kraemer et al., 2017). The gravity model (Erlander and Stewart, 1990) and the radiation model (Simini et al., 2012) are two models of human mobility that have been widely used in infectious disease modelling. The gravity model, originally formulated for transportation analysis (Erlander and Stewart, 1990) and later modified for infectious disease modelling, describes the number of individuals travelling from subpopulation i to subpopulation j as

$$T_{ij} \propto \frac{N_i^a N_j^b}{d_{ij}^c},$$

where N_i, N_j are the size of subpopulation i and j, respectively, d_{ij} is the distance between the two subpopulations, and a, b, c are constants. The parameter-free radiation model (Simini et al., 2012) and variants thereof (Yan et al., 2014; Kang et al., 2015) offer

alternative models for human mobility that only requires the spatial distribution of the population to estimate coupling. The standard radiation model describes the number of individuals travelling from subpopulation i to subpopulation j as

$$T_{ij} = \frac{N_i N_j}{(N_i + s_{ij})(N_i + N_j + s_{ij})} \sum_{k \neq i} T_{ik},$$

where s_{ij} is the population size in a circle of radius d_{ij} centered at subpopulation i.

Both the gravity and radiation model have been shown to capture observed commuting patterns at different spatial scales in Europe and North America (Balcan et al., 2009; Simini et al., 2012; Tizzoni et al., 2014; Yan et al., 2014). However, comparisons between these models and mobile call data records show that they fail to fully describe human mobility outside of high-income countries, such as in Sub-Saharan Africa (Wesolowski et al., 2015). These results reinforce observations by Horby et al. (2011) and Kiti et al. (2016) that contact patterns and household structure can vary between countries.

Non-spatial metapopulation models face other, distinct, challenges. For age-structured models, the transmission matrix is often based upon diary-based records of interactions (Mossong et al., 2008; Danon et al., 2013; Read et al., 2014). Diary-based methods face their own set of challenges, such as response bias, poor compliance, recall bias, and approximation of number and/or duration of contacts, especially for large numbers of contacts (Danon et al., 2013). For risk-structured models, it may be challenging to identify and interview hard-to-reach groups, such as intravenous drug users, men who have sex with men, or sex workers (Raymond and McFarland, 2009; Schneider et al., 2013).

Unlike social contact or mobility data, long-term data on disease incidence is often more widely available (Olsen and Schaffer, 1990; Grenfell and Harwood, 1997). In many countries, health professionals (broadly defined) are required to notify public health agencies about new cases of certain diseases. If this data can be split between the P subpopulations of interest, then we can measure the $\frac{1}{2}P(P-1)$ correlations between infection prevalence in each of the subpopulations. Given some relationship between the correlation and strength of interaction between subpopulations, we may be able to infer the coupling from disease prevalence data alone.

1.4 Discussion of key assumptions of infectious disease modelling

There are a number of key assumptions that are relatively standard in infectious disease modelling. These assumptions are often unrealistic for describing individual-level behaviour, but such models still show the correct population-level behaviour and are significantly easier to analyse and understand than their more complex alternatives. In this section we discuss the motivation and impact of these standard assumptions (although in general they are not challenged in this thesis).

The first key assumption is that model parameters are held constant through time. For the epidemic parameters (particularly the transmission rate β, but also, to a lesser extent, the recovery rate γ), this means that we assume there is no change due to intervention and control, or due to short- or long-term environmental changes, such as seasonality or long-term climate changes. Of these factors, seasonality is most often (and most easily) incorporated into models of infectious diseases, namely for those with recurent epidemics such as seasonal influenza or measles, by allowing the the transmission parameter to be a function of time (Keeling and Rohani, 2008); however, we do not consider seasonality in this thesis. By assuming that the demographic parameters (per capita birth rate ν and death rate μ) are constant, we assume no changes to the population growth rate or other economic or social factors that might affect this, including population fertility, life expectancy, living standards, and climate.

The second assumption is of homogeneous mixing, that is, that individuals mix and make contact randomly. Although this is an oversimplification at an individual level (we discuss this in detail in Section 1.3), this assumption confers two advantages. Firstly, the resulting model accurately captures the population-level dynamics in large populations. Secondly, assuming homogeneous mixing improves mathematical tractability of the model, since we do not need to understand or describe more complex patterns of interaction.

In stochastic formulations of infectious disease models, we also often assume that events occur at the points of independent Poisson processes with rates defined by the current state of the system. This means that the waiting time between, say, recovery events is exponentially distributed (or, equivalently, that the expected duration of the infectious period is exponentially distributed). At an individual level this is sometimes unsatisfactory: for example, the length of the infectious period will be typically clustered

around the mean and bounded above, in contrast to the exponential distribution which has no upper bound. However, as with the homogeneous mixing assumption, assuming exponential waiting times improves the mathematical tractability of the model. This key assumption underpins the use of Markov processes in infectious disease modelling (see Section) and consequently the use of Gillespie's algorithm and the τ-leaping algorithm to simulate Markov processes (see Section 1.2.4)

Chapter 2

Correlations between stochastic endemic infection in two interacting subpopulations

2.1 Introduction

In this chapter we derive an approximation for the correlation between the level of infection in two interacting populations as a function of the relative transmission rates, or the coupling, between them; this improves upon the results of Keeling and Rohani (2002) and corrects an error in the parametrisation of the original approximation. Using a multivariate normal moment closure approximation we derive this approximation for the simple case of two identical populations and provide conditions under which we expect this result to hold. We also numerically evaluate our model and compare our analytic approximation to stochastic simulations of the epidemic process.

2.2 A stochastic endemic infection model for two identical interacting populations

2.2.1 A simple endemic infection model

We begin by introducing the notation for a simple stochastic SIR model, with births, deaths, transmission and recovery. At any time $t \in [0, \infty)$, individuals are in one of three states: susceptible, infected or recovered.

A given susceptible individual meets infected individuals, and so themselves becomes infected, at rate $\beta > 0$. Susceptible individuals can also succumb to infection independent of contact with infected individuals in the populations; this occurs at rate $\epsilon > 0$, the external import rate. Infected individuals recover from infection at rate $\gamma > 0$, and individuals die at rate $\mu > 0$, independent of infection status. We assume that a death is immediately followed by the birth of a susceptible individual, and hence the total population size remains constant. The basic reproductive ratio, R_0, for this process is $R_0 = \beta/(\gamma + \mu)$. Let $S(t), I(t), R(t) \in \{0, 1, 2, \ldots\}$ denote the number of susceptible, infected and recovered individuals, respectively, at time $t \geq 0$. If we let the (constant) population size be equal to N, we can reduce the dimensionality of the system by setting $R(t) = N - S(t) - I(t)$. The Kolmogorov forward equations for this process are given by

$$\begin{aligned}\frac{dp_t(s,i)}{dt} = &\left(\frac{\beta}{N}(s+1)(i-1) + \epsilon(s+1)\right) p_t(s+1, i-1) + \gamma(i+1)p_t(s, i+1) \\ &+ \mu(i+1)p_t(s-1, i+1) + \mu(N-(s-1)-i)p_t(s-1, i) \\ &- \left(\frac{\beta}{N}si + \epsilon s + \gamma i + \mu i + \mu(N-s-i)\right) p_t(s, i),\end{aligned} \tag{2.1}$$

where $p_t(s, i)$ is the probability that there are s susceptible individuals and i infectious individuals in the population at time t.

2.2.2 An endemic infection model for coupled populations

Consider a pair of identical populations of size N. We assume the populations are the same size for analytical tractability; we discuss the effect of relaxing this assumption in Section 2.3. Furthermore, we assume that both populations exhibit the same population

dynamics as the simple stochastic epidemic model described in Section 2.2.1; however, we now assume that a proportion $\sigma \in [0,1]$ of an individual's contacts are with individuals in the other population. In this way, σ describes the interaction, or 'coupling', between the two populations, and the force of infection in each population depends on the number of infected individuals in both populations. Changing σ does not change the basic reproductive ratio in this model, but simply determines the distribution of secondary cases between the two populations.

We now let $S_j(t), I_j(t), R_j(t) \in \{0,1,2,\ldots\}$ denote the number of susceptible, infected and recovered individuals, respectively, in population $j = 1,2$ at time $t \geq 0$; and again insist that population sizes remain constant: $N = S_j(t) + I_j(t) + R_j(t), \forall t \geq 0, j = 1,2$. The transition rates for the resulting four-dimensional Markov process from state (s_1, i_1, s_2, i_2) at time t are summarised in Table 2.1.

We can also write down the the Kolmogorov forward equations for this process. Let $p_t(s_1, i_1, s_2, i_2) = \mathbb{P}((S_1(t), I_1(t), S_2(t), I_2(t)) = (s_1, i_2, s_2, i_2))$. The Kolmogorov forward equation for the stochastic epidemic model for two identical coupled populations is given by:

$$\begin{aligned}
\frac{dp_t(s_1,i_1,s_2,i_2)}{dt} = &\left(\frac{\beta}{N}(s_1+1)[(1-\sigma)(i_1-1)+\sigma i_2]+\epsilon(s_1+1)\right)p_t(s_1+1,i_1-1,s_2,i_2)\\
&+\left(\frac{\beta}{N}(s_2+1)[\sigma i_1+(1-\sigma)(i_2-1)]+\epsilon(s_2+1)\right)p_t(s_1,i_1,s_2+1,i_2-1)\\
&+\gamma(i_1+1)p_t(s_1,i_1+1,s_2,i_2)+\gamma(i_2+1)p_t(s_1,i_1,s_2,i_2+1)\\
&+\mu(i_1+1)p_t(s_1-1,i_1+1,s_2,i_2)+\mu(i_2+1)p_t(s_1,i_1,s_2-1,i_2+1)\\
&+\mu(N-(s_1-1)-i_1)p_t(s_1-1,i_1,s_2,i_2)\\
&+\mu(N-(s_2-1)-i_2)p_t(s_1,i_1,s_2-1,i_2)\\
&-\left(\frac{\beta}{N}s_1[(1-\sigma)i_1+\sigma i_2]+\epsilon s_1+\gamma i_1+\mu i_1+\mu(N-s_1-i_1)\right.\\
&\left.+\frac{\beta}{N}s_2[\sigma i_1+(1-\sigma)i_2]+\epsilon s_2+\gamma i_2+\mu i_2+\mu(N-s_2-i_2)\right)p_t(s_1,i_1,s_2,i_2).
\end{aligned} \tag{2.2}$$

Population	Event	Transition	Rate
$j, k \in \{1, 2\}$,	Infection	$s_j \to s_j - 1, i_j \to i_j + 1$	$\beta s_j[(1-\sigma)i_j + \sigma i_k]/N + \epsilon s_j$
$k \neq j$	Recovery	$i_j \to i_j - 1, r_j \to r_j + 1$	γi_j
	Death of infected	$s_j \to s_j + 1, i_1 \to i_j - 1$	μi_j
	Death of recovered	$s_j \to s_j + 1, r_j \to r_j - 1$	$\mu(N - s_j - i_j)$

Table 2.1. A summary of the transition rates of the four-dimensional Markov process epidemic model $\{(S_1(t), I_1(t), S_2(t), I_2(t)) : t \geq 0\}$ from state (s_1, i_1, s_2, i_2) with birth/death rate $\mu > 0$, contact rate $\beta > 0$, external import rate $\epsilon > 0$, recovery rate $\gamma > 0$ and coupling $\sigma \in [0, 1]$.

2.2.3 Dynamics of first- and second-order moments

Theoretical

As exact analysis of the coupled stochastic epidemic model is mathematically intractable, we consider the approximate behaviour of the first- and second-order central moments of the process. For the coupled stochastic epidemic model there are eight distinct first- and second-order central moments, five of which are 'within-population' and three of which are 'between-population'. Since the two populations are identical, there are symmetries within the system that can be exploited: for example, $\mathbb{E}[S_1] = \mathbb{E}[S_2]$ and $Var(S_1) = Var(S_2)$, and similarly for other central moments. We denote the within-population central moments by

$$\begin{aligned} \bar{S} &= \mathbb{E}[S_1] = \mathbb{E}[S_2] \\ \bar{I} &= \mathbb{E}[I_1] = \mathbb{E}[I_2] \\ C_{SS} &= Var(S_1) = Var(S_2) \\ C_{II} &= Var(I_1) = Var(I_2) \\ C_{SI} &= Cov(S_1, I_1) = Cov(S_2, I_2) \end{aligned}$$

and denote the between-population central moments by

$$\begin{aligned} \hat{C}_{SS} &= Cov(S_1, S_2) \\ \hat{C}_{II} &= Cov(I_1, I_2) \\ \hat{C}_{SI} &= Cov(S_1, I_2) = Cov(S_2, I_1). \end{aligned}$$

We can write down an ODE for each of the eight first- and second-order central moments using the Kolmogorov forward equation, using the method outlined in Section 1.2.2; in summary, the ODE for $\mathbb{E}[f(X)]$ is given by:

$$\frac{d\mathbb{E}[f(X)]}{dt} = \mathbb{E}\left[\sum_{\text{events}} \text{rate of event} \times \text{change in } f(X) \text{ due to event}\right]. \tag{2.3}$$

Due to the non-linearity of the infection term in the model, the ODE for an n-th-order moment will depend on one or more $(n+1)$-th-order moments. To fully define the system of ODEs we would therefore have to write down an infinite set of equations. To circumvent this problem we use a moment closure approximation, which truncates this set of equations at some order. Here, we make a second-order moment closure approximation, which assumes that third- and higher-order cumulants are equal to zero. In this way, third-order moments can be written in terms of the mean and covariance. This is equivalent to assuming that the random variable has a multivariate normal (MVN) distribution (Whittle, 1957) and so we refer to this approximation as a second-order MVN moment closure approximation. The resulting set of eight ODEs and their derivation can be found in Appendix A. Note that a first-order moment closure approximation assumes that second- and higher-order cumulants are equal to zero; this approximation returns the standard set of ODEs for the SIR-model, which describe the stochastic process in the large-population limit.

Dynamics of a measles-like disease in the UK

For the majority of the numerical analysis, we utilise parameters for a highly-transmissible measles-like endemic disease in the UK (Anderson and May, 1992), although we note that a full model of measles requires both seasonality (Earn et al., 2000; Rohani et al., 2002; Grenfell and Bolker, 1995) and age-structure (Schenzle, 1984; Keeling and Grenfell, 1997; Bolker, 1993). We consider two identical populations of size $N = 10^5$ where $R_0 = 17$, $\gamma^{-1} = 13$ days, $\mu = 5.5 \times 10^{-5}$ days^{-1} and $\epsilon = 5.5 \times 10^{-5}$ days^{-1}. The numerical integration of ODEs is performed using the MATLAB ode45 solver with a relative error tolerance of 10^{-5}.

Figure 2.1 shows the equilibrium values of the first-order central moments $\bar{S}^*$ and $\bar{I}^*$ and second-order central moments C^*_{II} and $\hat{C}^*_{II}$ for a measles-like endemic disease in the UK as the coupling parameter σ is varied between 0 and 1. These results are obtained by numerical integration of 8-dimensional ODE system given in Appendix A,

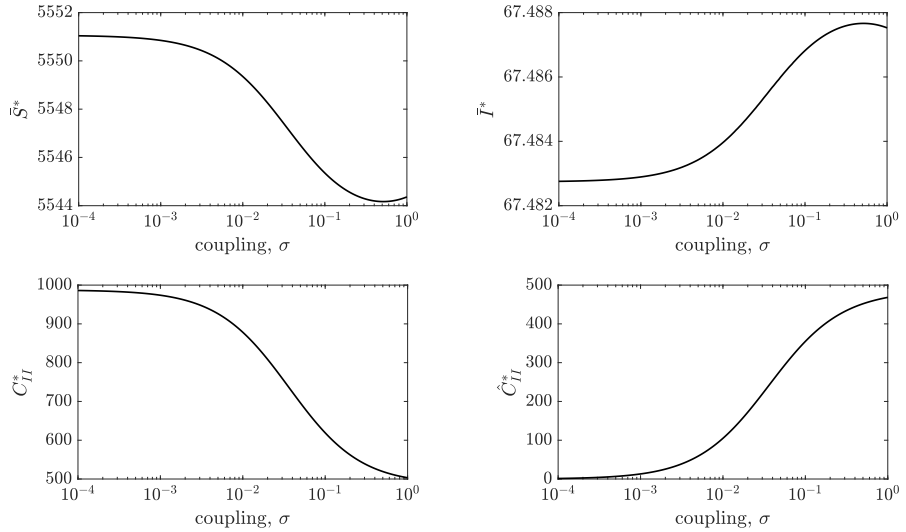


Figure 2.1. The effect of the coupling, σ, on key mean variables $\bar{S}^*, \bar{I}^*, C^*_{II}$ and $\hat{C}^*_{II}$ for a measles-like endemic disease in the UK ($N = 10^5, \mu = 5.5 \times 10^{-5}, R_0 = 17, \gamma^{-1} = 13$ and $\epsilon = 5.5 \times 10^{-5}$), calculated from the ODEs given in Appendix A.

and therefore only depend on the MVN moment closure approximation. We note that all curves broadly show a sigmoidal pattern (although $\bar{S}^*$ has a minimum and $\bar{I}^*$ a maximum at $\sigma = 0.5$), with $\bar{S}^*$ and C^*_{II} decreasing with the coupling and $\bar{I}^*$ and $\hat{C}^*_{II}$ increasing with the coupling.

2.2.4 Approximation for the correlation between populations

We derive a theoretical approximation for the correlation at endemic equilibrium between the number of infected individuals in population 1 and the number of infected individuals in population 2 as a function of the coupling, σ. We define the correlation between the number of infected individuals in each population at endemic equilibrium as:

$$\rho = \frac{Cov(I_1, I_2)}{\sqrt{Var(I_1)Var(I_2)}},$$

which, in the case of two identical populations where the variances are equal and using our earlier notation, simplifies to:

$$\rho = \frac{\hat{C}^*_{II}}{C^*_{II}},$$

where X^* denotes the quantity X at endemic equilibrium.

For two identical populations we find that we can write the correlation as a sigmoidal function plus a correction term that is often relatively small:

$$\rho = \frac{\sigma}{\xi + \sigma} - \Delta, \tag{2.4}$$

where

$$\xi = \frac{N(\gamma + \mu) - \beta \bar{S}^*}{\beta \bar{S}^*} \tag{2.5}$$

and

$$\Delta = \frac{\left(\beta \bar{I}^* + N\epsilon\right) \frac{\hat{C}^*_{SI}}{C^*_{II}}}{\beta(1-\sigma)\bar{S}^* - N(\gamma + \mu)}. \tag{2.6}$$

To derive this result, we use the moment equation for $\hat{C}_{II}$ derived in Appendix A:

$$\frac{d\hat{C}_{II}}{dt} = 2\frac{\beta}{N}\sigma \bar{S} C_{II} + 2\left(\frac{\beta}{N}(1-\sigma)\bar{S} - \gamma - \mu\right)\hat{C}_{II} + 2\left(\frac{\beta}{N}\bar{I} + \epsilon\right)\hat{C}_{SI}. \tag{2.7}$$

At equilibrium $d\hat{C}_{II}/dt = 0$ and if we divide by $2C^*_{II}/N$, then

$$0 = \beta\sigma\bar{S}^* + \left(\beta(1-\sigma)\bar{S}^* - N(\gamma+\mu)\right)\rho + \left(\beta\bar{I}^* + N\epsilon\right)\frac{\hat{C}^*_{SI}}{C^*_{II}}, \tag{2.8}$$

and hence we have the following approximation for the correlation that we will henceforth refer to as the MVN correlation:

$$\begin{aligned}
\rho &= \frac{-\beta\sigma\bar{S}^*}{\beta(1-\sigma)\bar{S}^* - N(\gamma+\mu)} - \frac{\left(\beta\bar{I}^* + N\epsilon\right)\frac{\hat{C}^*_{SI}}{C^*_{II}}}{\beta(1-\sigma)\bar{S}^* - N(\gamma+\mu)} \\
&= \frac{\sigma}{\left(\frac{N(\gamma+\mu)-\beta\bar{S}^*}{\beta\bar{S}^*}\right) + \sigma} - \Delta \\
&= \frac{\sigma}{\xi + \sigma} - \Delta.
\end{aligned} \tag{2.9}$$

Moreover, if we can show that $\Delta \ll 1$ then we have the following simplified approximation for the correlation

$$\rho \approx \frac{\sigma}{\xi + \sigma}. \tag{2.10}$$

Alternative parametrisation of ξ

We can also derive an alternative approximate expression for ξ that is independent of $\bar{S}^*$, hence eliminating the need to find the equilibrium of the 8-dimensional ODE system. By ignoring the effects of both imports and correlations and taking the large population limit, we can find an approximation to S^*, which leads to the following expression:

$$\xi \approx \xi' = \frac{\epsilon(\gamma+\mu)}{\mu(\beta-\gamma-\mu)} = \frac{\epsilon}{\mu(R_0-1)}. \tag{2.11}$$

To derive this result, we begin with the ODE for $\bar{I}$ given in Appendix A; in the case where $\sigma = 0$ then we have

$$\frac{d\bar{I}}{dt} = \frac{\beta}{N}(\bar{S}\bar{I} + C_{SI}) + \epsilon\bar{S} - (\gamma+\mu)\bar{I}. \tag{2.12}$$

At equilibrium, $d\bar{I}/dt = 0$ and so

$$\frac{\beta}{N}(\bar{S}^*\bar{I}^* + C^*_{SI}) + \epsilon\bar{S}^* - (\gamma+\mu)\bar{I}^* = 0 \tag{2.13}$$

$$\iff \beta\bar{S}^* + \frac{\beta C^*_{SI}}{\bar{I}^*} + \frac{N\epsilon\bar{S}^*}{\bar{I}^*} - N(\gamma+\mu) = 0 \tag{2.14}$$

$$\iff \xi = \frac{N(\gamma+\mu) - \beta\bar{S}^*}{\beta\bar{S}^*} = \frac{N\epsilon}{\beta\bar{I}^*} + \frac{C^*_{SI}}{\bar{S}^*\bar{I}^*}. \tag{2.15}$$

Note that since $\bar{S}^*, \bar{I}^*, C^*_{SI} = O(N)$ then $C^*_{SI}/\bar{S}^*\bar{I}^* = O(1/N)$.

In the large-population limit with no external imports ($\epsilon = 0$)

$$\frac{d\bar{S}_0}{dt} = \mu N - \frac{\beta}{N}\bar{S}_0\bar{I}_0 - \mu\bar{S}_0 \tag{2.16}$$

$$\frac{d\bar{I}_0}{dt} = \frac{\beta}{N}\bar{S}_0\bar{I}_0 - (\gamma+\mu)\bar{I}_0, \tag{2.17}$$

and since $\epsilon \ll 1$, we make the simplifying assumption that $\epsilon^2 \approx 0$. At equilibrium, then $\bar{I}^*_0 = N\mu(\beta-\gamma-\mu)/(\beta(\gamma+\mu)) = N\mu(R_0-1)/\beta$ and we write $\bar{I}^* = \bar{I}^*_0 + O(N\epsilon, 1)$.

Therefore, in the large-population limit we have

$$\xi = \frac{N\epsilon}{\beta\bar{I}^*} + \frac{C^*_{SI}}{\bar{S}^*\bar{I}^*} = \frac{N\epsilon}{\beta(\bar{I}^*_0 + O(N\epsilon, 1))} + O(1/N) \tag{2.18}$$

$$= \frac{N\epsilon}{\beta\bar{I}^*_0} \frac{1}{1 + \frac{O(N\epsilon,1)}{\bar{I}^*_0}} \tag{2.19}$$

$$= \frac{N\epsilon}{\beta\bar{I}^*_0} \left(1 - \frac{O(N\epsilon, 1)}{\bar{I}^*_0}\right) \tag{2.20}$$

$$= \frac{N\epsilon}{\beta\bar{I}^*_0}(1 - O(\epsilon, 1/N)) \tag{2.21}$$

$$= \frac{\epsilon(\gamma + \mu)}{\mu(\beta - \gamma - \mu)}(1 - O(\epsilon, 1/N)) \tag{2.22}$$

$$= \frac{\epsilon}{\mu(R_0 - 1)}. \tag{2.23}$$

This parametrisation of ξ is preferable to the original (Equation (2.5)) as it removes the need to estimate the number of susceptible individuals in the population at endemic equilibrium, either from data or through simulation. In addition, this alternative parametrisation provides intuition into how the epidemic parameters directly impact the correlation. We can see that as R_0 increases then the correlation also increases. Conversely, as the external import rate ϵ increases, then the correlation decreases: as ϵ increases then external infections mask the effect of the between-population infections. Given the appeal of the simpler form of Equation (2.11), in the work that follows we evaluate the approximation of the correlation ρ by the sigmoidal function $\sigma/(\xi' + \sigma)$.

2.2.5 Evaluating our approximation for the correlation

The MVN moment closure approximation holds in the large-population limit (i.e $N \to \infty$) and assumes that the distribution of states is a multivariate normal distribution; this follows from the results of (Kurtz, 1970, 1971), which show that a stochastic process can be approximated by a deterministic processes in the large population limit. Further error in approximation comes from assuming that $\Delta \ll 1$ and ξ is constant and equal to ξ'. In the following section, our aim is to understand whether our approximation (Equation (2.10)) and expression for ξ' (Equation (2.11)) are generic to a wider range of assumptions and parameters.

Using the parameters for a measles-like disease in the UK, we solve the underlying ODEs and hence check numerically that Δ is small and calculate ξ. The absolute error

introduced into our approximation by assuming that $\Delta \ll 1$ is given by Δ; the error relative to the correlation ρ is given by Δ/ρ. The absolute error introduced into our approximation by assuming that ξ is constant (Equation 2.11) is $[\rho - \sigma/(\xi' + \sigma) + \Delta]$, or equivalently $[\sigma/(\xi + \sigma) - \sigma/(\xi' + \sigma)]$; the error relative to the correlation ρ is given by $[\sigma/(\xi + \sigma) - \sigma/(\xi' + \sigma)]/\rho$. In this chapter we take 0.1 as a threshold for the absolute error and 0.25 as a threshold for the error relative to the correlation ρ. If the absolute or relative error exceed 0.1 or 0.25 respectively, then we say that the approximation fails.

For the parameters for a measles-like disease in the UK, we compare the MVN correlation ρ (Equation (2.9)) and our approximation $\sigma/(\xi' + \sigma)$, $\xi' = 0.0625$, to the results of full stochastic simulations (Figure 2.2). We simulate the stochastic process over a 200 year period using the Gillespie algorithm, with a burn-in period of 50 years, and generate 1000 realisations of the process for each value of σ. The correlation is calculated as a time-weighted Pearson correlation coefficient for $50 < t \leq 200$ years. From this comparison we draw three conclusions. Firstly, all three correlations follow a sigmoidal relationship increasing from zero for low coupling to a value close to one when the coupling is largest- although we note that values of $\sigma > 0.5$ do not match with our idealised view of a metapopulation in which within-population transmission is larger than between-population transmission. Secondly, the remarkably close agreement between ρ and the simulation results, suggest our use of the MVN moment closure approximation is justified. Finally, $\sigma/(\xi' + \sigma)$ is a reasonable approximation for the MVN correlation ρ as the difference between the two curves is small.

In Figure 2.3, we evaluate the two main sources of error in our approximation (Equation (2.10)), introduced by assuming that $\Delta \ll 1$ and that ξ is constant and equal to ξ'. For measles-like parameters, Δ is small in both absolute and relative terms (green lines in Figure 2.3); importantly, Δ never exceeds our chosen thresholds of 0.1 and 0.25 for the absolute and relative error respectively, and has a diminishing impact on the correlation when the coupling σ exceeds 0.065. The error introduced into the approximation by assuming ξ is constant is given by $[\rho - \sigma/(0.0625 + \sigma) + \Delta] = [\sigma/(\xi + \sigma) - \sigma/(0.0625 + \sigma)]$ (yellow lines in Figure 2.3). Again, we observe that this error is well within our chosen thresholds of 0.1 and 0.25 for the absolute and relative error respectively. Moreover, this error is approximately one order of magnitude smaller than Δ, which tells us that for measles-like parameters, the main source of error is our approximation is due to assuming that $\Delta \ll 1$. Overall, these findings suggest that our simple approximation (Equation (2.10)) should hold for these parameters across the entire range of coupling values.

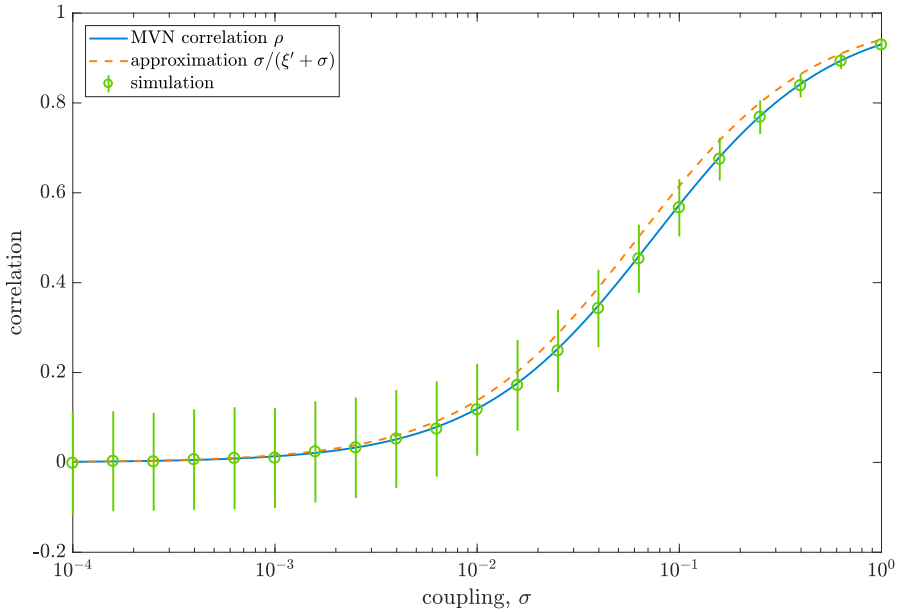


Figure 2.2. Comparing the MVN correlation ρ and our approximation $\sigma/(\xi' + \sigma)$ to stochastic simulations, for a measles-like endemic disease in the UK ($N = 10^5, \mu = 5.5 \times 10^{-5}, R_0 = 17, \gamma^{-1} = 13$ and $\epsilon = 5.5 \times 10^{-5}$; $\xi' = 0.0625$). We generate 1000 realisations of the process for each value of σ and calculate the correlation as a time-weighted Pearson correlation coefficient for $50 < t \leq 200$ years; error bars represent ± 2 standard deviations.

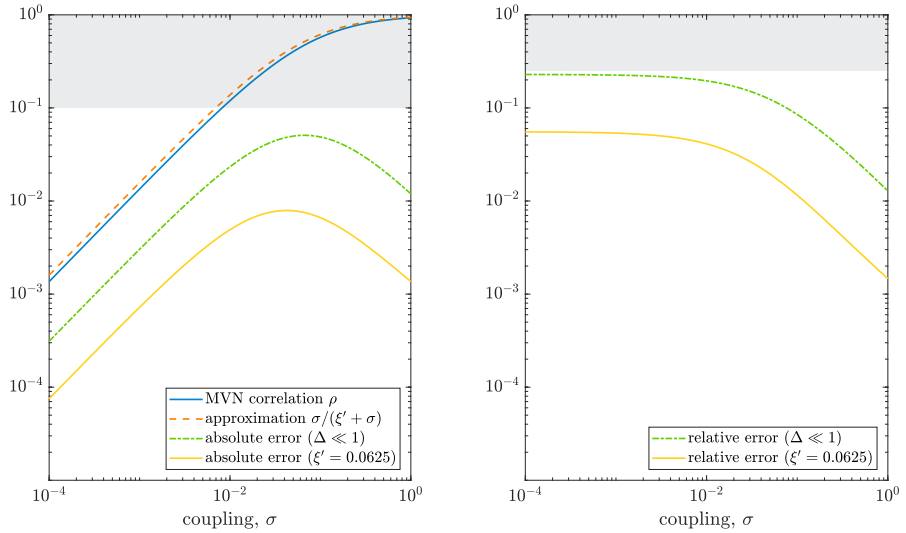


Figure 2.3. Evaluating the sources of error in our approximation for a measles-like endemic disease in the UK ($N = 10^5, \mu = 5.5 \times 10^{-5}, R_0 = 17, \gamma^{-1} = 13$ and $\epsilon = 5.5 \times 10^{-5}$), taking 0.1 and 0.25 as thresholds for the absolute (left) and relative (right) error, respectively. We compare the MVN correlation ρ and our approximation $\sigma/(\xi'+\sigma)$ to the two sources of error in our approximation: assuming $\Delta \ll 1$, and assuming that ξ is constant and equal to $\xi' = \epsilon/(\mu(R_0 - 1)) = 0.0625$.

2.2.6 Sensitivity analysis

Smaller population sizes

Whilst the MVN moment closure approximation holds in the large population limit (Kurtz, 1970, 1971), we may often be interested in much smaller populations where the impact of stochasticity is more pronounced. For $N = 10^2, 10^3, 10^4, 10^5$ we compare our approximation for the correlation to stochastic simulations. We generate 1000 realisations of each (N, σ) pair and calculate the correlation between the two populations. Since $\xi' = \epsilon/(\mu(R_0 - 1))$ is independent of N, we take $\xi' = 0.0625$ for all N.

In Figure 2.4 we compare the stochastic simulations for each of $N = 10^2, 10^3, 10^4, 10^5$ to our approximation $\sigma/(0.0625+\sigma)$. We find that, for a given σ, decreasing the population size leads to weaker correlations; equivalently, this means that in smaller populations stronger coupling is required to achieve the same level of correlation. This is because in small populations extinction events are more frequent and so the dynamics of each population are characterised by periods of zero infection and (random) reinfection events; this behaviour acts to reduce the correlation between the two populations. Despite this, we find even for $N = 10^3$ the correlation between the two populations is well approximated by $\sigma/(\xi'+\sigma)$; only at very small population sizes, $N = 100$, is our approximation a poor estimate of the correlation.

Although it is simpler to take $\xi = \xi'$, this value can also be calculated as $\xi = (N(\gamma + \mu) - \beta\bar{S}^*)/\beta\bar{S}^*$, for some specific value of $\bar{S}^*$. This method requires the numerical integration of the ODEs given in Appendix A; however, we find that for $N \lesssim 10^{4.2} \approx 16,000$ the numerical solution to the system of ODEs "blows up". Therefore, we cannot use this method for calculating ξ to parametrise our approximation in smaller populations. This phenomenon occurs since we assume that the distribution of states follows a multivariate normal distribution and at low levels of infection this leads to a significant proportion of the distribution being negative. For example, for $N = 10^5$ then $\bar{I}^* \approx 67$, but for $N \lesssim 10^{4.2}$ then $\bar{I} \lesssim 10.6$. Zero infectious cases should act as a boundary for the distribution, and hence as $\bar{I}$ decreases, the multivariate normal assumption breaks down.

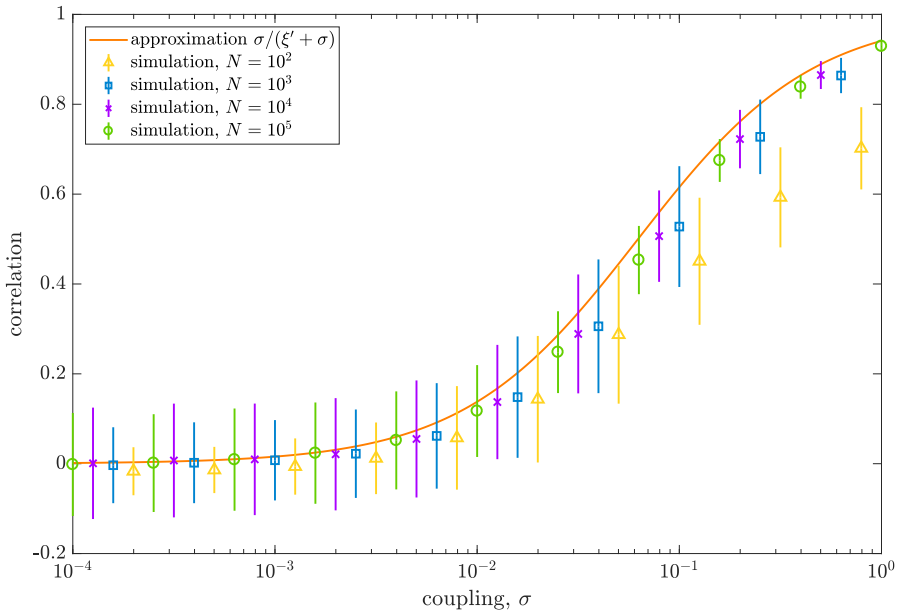


Figure 2.4. Comparing our approximation $\sigma/(\xi' + \sigma)$, $\xi' = 0.0625$, to stochastic simulations for a measles-like endemic disease in the UK ($\mu = 5.5 \times 10^{-5}, R_0 = 17, \gamma^{-1} = 13$ and $\epsilon = 5.5 \times 10^{-5}$) and $N = 10^2, 10^3, 10^4, 10^5$. We simulate the stochastic process over a 200 year period using the Gillespie algorithm, with a burn-in period of 50 years, and generate 1000 realisations of the process for each of (N, σ) pair. The correlation is calculated as a time-weighted Pearson correlation coefficient for $50 < t \leq 200$; error bars represent ± 2 standard deviations.

Parameter sensitivity analysis

We perform a brief parameter sensitivity analysis to understand how the correlation between the number of infected individuals in the two populations is affected by the values of the epidemiological parameters. In the first half of the analysis, we use the parameters for a measles-like disease in the UK ($N = 10^5, \mu = 5.5\times10^{-5}, R_0 = 17, \gamma^{-1} = 13$ and $\epsilon = 5.5\times10^{-5}$) and independently change the value of each of the four parameters μ, β, ϵ and γ. We show the impact of these epidemiological parameters on $\xi' = \epsilon/(\mu(R_0 - 1))$. For each set of epidemiological parameter values we also calculate Δ and compare $\xi \sim \xi'$ across all values of coupling σ to determine the range of parameter values for which our approximation holds, that is, for which the absolute and relative errors are within our chosen thresholds of 0.1 and 0.25 respectively for all coupling values.

We find that the values of each of the four key parameters have a profound impact on the correlation between the number of infected individuals in the two populations, but that our approximation holds for a wide range of realistic values (Figure 2.5). The correlation increases with the birth rate, μ, the basic reproductive ratio, $R_0 = \beta/(\gamma+\mu)$, varied by changing β, and the mean infectious period, γ^{-1}; increases in the external import rate, ϵ, lead to a decrease in the correlation. The exact region in which our approximation fails is a complex trade-off between all parameters; however, for all four parameters the approximation fails (that is, either the absolute or relative error exceeds our chosen thresholds) as ξ' becomes smaller; failures occur for $\mu \gtrsim 5.62\times10^{-5}$, $\beta \gtrsim 1.35$ ($R_0 \gtrsim 17.5$), $\epsilon \lesssim 4.61 \times 10^{-5}$ and $\gamma^{-1} \gtrsim 13$. This failure mode is due to the growing importance of the correction term Δ relative to our approximation $\sigma/(\xi' + \sigma)$.

Other childhood diseases

We also compare the MVN correlation ρ and our approximation $\sigma/(\xi' + \sigma)$ using parameter values for six other infectious diseases in the UK: mumps, rubella, chickenpox, whooping cough (Anderson and May, 1992), smallpox (Keeling and Rohani, 2008) and influenza (Cauchemez et al., 2004; Biggerstaff et al., 2014) (Figure 2.6) Disease-specific parameters are given in Table 2.2; for all diseases, we consider two identical populations of size $N = 10^5$ where $\mu = 5.5 \times 10^{-5}$ days^{-1} and $\epsilon = 5.5 \times 10^{-5}$ days^{-1}.

Interestingly, we observe that our approximation overestimates the correlation for diseases with a high R_0 (e.g. whooping cough) and underestimates the correlation for diseases with a low R_0 (e.g. influenza). We attribute this to the differential action of

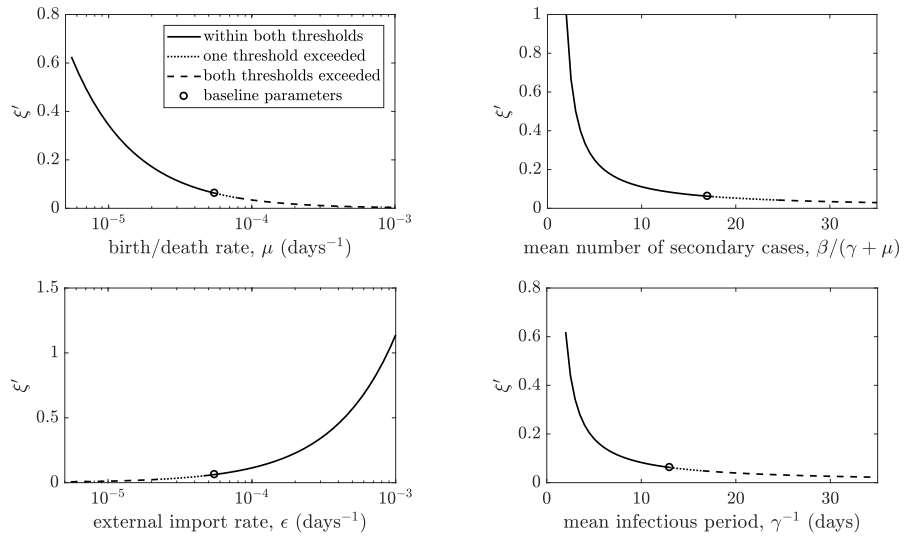


Figure 2.5. Sensitivity analysis for our approximation to the epidemic parameters. The parameter ξ in our approximation is calculated as $\xi' = \epsilon/(\mu(R_0 - 1))$; our approximation holds if both the absolute and relative error is within our chosen thresholds of 0.1 and 0.25 respectively (represented by a solid line). The approximation fails due to one or both of the absolute and relative errors exceeding our chosen thresholds (dotted and dashed lines respectively). This analysis is performed for each of the four epidemiological parameters μ, β, ϵ and γ with baseline parameters $\mu = 5.5 \times 10^{-5}, R_0 = 17, \gamma^{-1} = 13$ and $\epsilon = 5.5 \times 10^{-5}$ (shown by a circle). $N = 10^5$ throughout.

Disease	Basic reproductive ratio R_0	Average infectious period γ^{-1} (days)	ξ'
Mumps	12	21	0.0909
Rubella	7	17	0.1667
Chickenpox	11	20	0.1
Whooping cough	17	22	0.0625
Smallpox	5	7	0.25
Influenza	2	4	1

Table 2.2. Epidemiological parameters for seven infectious diseases in the UK; across all diseases we take $N = 10^5, \mu = 5.5 \times 10^{-5}$ days^{-1} and $\epsilon = 5.5 \times 10^{-5}$ days^{-1}. We also give the value of the parameter $\xi' = \epsilon/(\mu(R_0 - 1))$ taken in our approximation for the correlation.

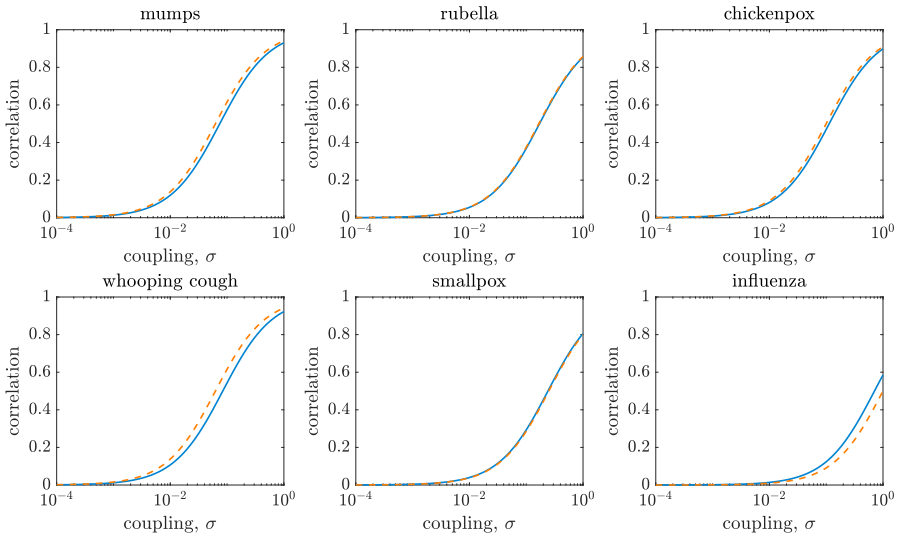


Figure 2.6. Comparing the MVN correlation ρ and our approximation $\sigma/(\xi' + \sigma)$ for parameters representing mumps, rubella, chickenpox, whooping cough, smallpox and influenza (parameter values are given in Table 2.2).

Δ and the approximation to ξ across epidemiological parameters. However, across all diseases the difference between ρ and our approximation is small, hence we can relate the phenomenological coupling parameter, σ, to the correlation between the number of infected individuals in two populations by $\rho = \sigma/(\xi' + \sigma)$.

In the second half of the analysis, we focus on the external import rate ϵ, as this is generally the most difficult parameter to estimate. For each of the epidemiological parameter sets representing mumps, rubella, chickenpox, whooping cough, smallpox and

influenza, and for each value of ϵ, we show $\xi' = \epsilon/(\mu(R_0-1))$. For each disease parameter set we also determine a range of values for ϵ for which the approximation holds, that is, for which the absolute and relative error are within our chosen thresholds of 0.1 and 0.25 respectively.

We find that the external import rate ϵ has a significant impact on the correlation (Figure 2.7). For all diseases we consider, increasing the external import rate leads to a higher value of ξ' and thus predicts a lower correlation for a given coupling strength: as the external import rate is increased, external infections mask the effect of the between-population infections. For measles, mumps, rubella, chickenpox, whooping cough and smallpox, the approximation fails (that is, either the absolute or relative error exceeds our chosen thresholds) as ϵ become smaller. We also observe that our approximation fails at lower values of ϵ in diseases with a lower R_0: for example, our approximation for whooping cough ($R_0 = 17$) fails for $\epsilon \lesssim 8.71 \times 10^{-5}$, whereas our approximation for rubella ($R_0 = 7$) fails for $\epsilon \lesssim 9.78 \times 10^{-6}$. However, for influenza, our approximation fails as ϵ becomes larger.

We compare ξ' and $\xi = (N(\gamma+\mu) - \beta\bar{S}^*)/\beta\bar{S}^*$ for each disease (Figure 2.8). Since ξ changes with σ, we choose here to plot $\min_\sigma \xi$; however, we could plot $\max_\sigma \xi$ and infer the same result. This analysis shows that for influenza, $\xi' = \epsilon/\mu(R_0 - 1))$ significantly overestimates ξ for large values of ϵ, so assuming that $xi = \xi'$ leads to a large error in our approximation. For the other diseases considered, the difference between ξ' and ξ is small and so we do not observe failure for large values of ϵ in the range of values that we consider (Figure 2.8); the value of ϵ would have to be unrealistically high to observe such an effect.

2.3 A stochastic endemic infection model for two non-identical interacting populations

If we relax the assumption that the two populations are of equal size, then we do not obtain the same simple relationship between the coupling and correlation. The model is less amenable to analytic methods and the resulting approximation now depends on $r = N_1/N_2$ and on the equilibrium values $\bar{S}_j^*$ and $Var(I_j), j = 1, 2$. To fully define the approximation we would need to estimate these values from data or through simulation.

Consider a pair of populations of size N_1 and N_2 respectively, where without loss of generality we assume $N_2 = rN_1, r > 1$. Both populations exhibit the same population

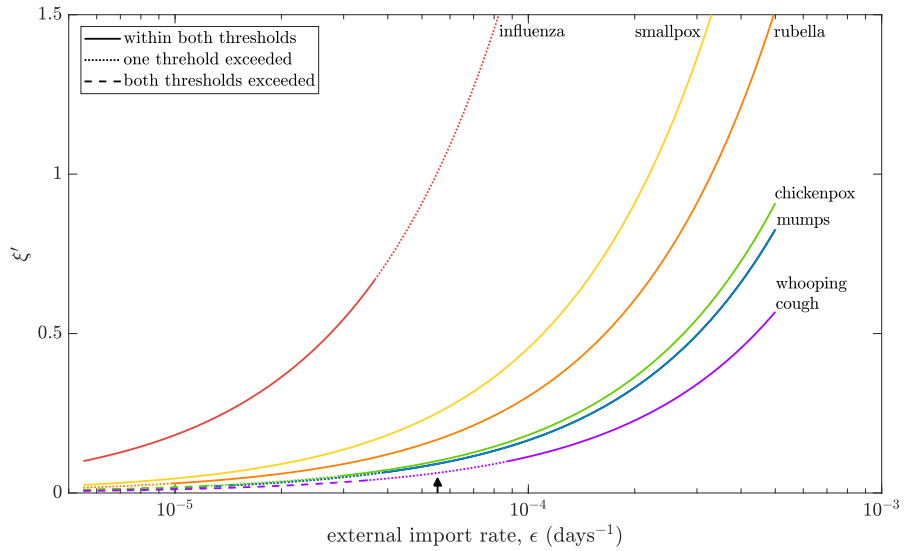


Figure 2.7. Sensitivity analysis for our approximation to the epidemic parameters. The parameter ξ in our approximation is calculated as $\xi' = \epsilon/(\mu(R_0 - 1))$; our approximation holds if both the absolute and relative error is within our chosen thresholds of 0.1 and 0.25 respectively (represented by a solid line). The approximation fails due to one or both of the absolute and relative errors exceeding our chosen thresholds (dotted and dashed lines respectively). This analysis is performed for the external import rate ϵ for the given diseases with baseline parameters given in Table 2.2 (shown by an arrow on the x-axis). $N = 10^5$ throughout.

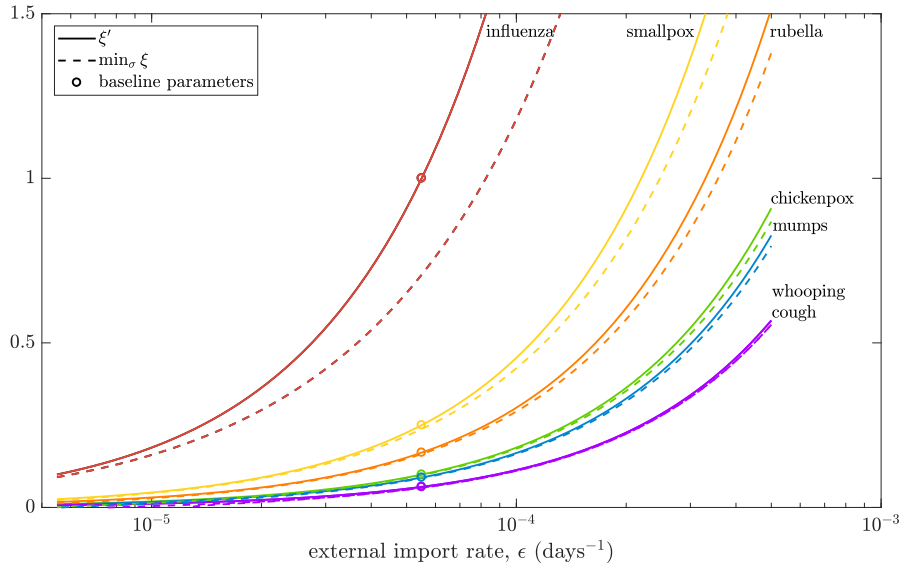


Figure 2.8. We assess the assumption that $\xi = \xi'$ in our approximation and compare ξ' and $\min_\sigma \xi$ for each of influenza, smallpox, rubella, chickenpox, mumps and whooping cough. We note that for influenza ξ' overestimates ξ significantly for large values of ϵ, which leads to a large error in our approximation.

dynamics as the simple stochastic epidemic model described in Section 2.1; however, we now assume that in population $j = 1, 2$ a proportion $\sigma_j \in [0, 1]$ of an individual's contacts are with individuals in the other population. To maintain a constant population size, then we assume that $(1 - \sigma_1)N_1 + \sigma_2 N_2 = N_1$, and so $\sigma_2 = \sigma_1/r$.

The correlation between the number of infected individuals in each population at endemic equilibrium as:

$$\rho = \frac{\hat{C}^*_{II}}{\sqrt{V_1^* V_2^*}},$$

where $V_j^* = Var(I_j), j = 1, 2$.

We find that we can write the correlation as a sigmoidal function plus a correction term:

$$\rho = \sqrt{\frac{V_2^*}{V_1^*}} \frac{\sigma_1}{\frac{N_2}{N_1}(\xi_1 + \sigma_1) + \frac{S_2^*}{S_1^*}(\xi_2 + \sigma_2)} + \sqrt{\frac{V_1^*}{V_2^*}} \frac{\sigma_2}{\frac{N_1}{N_2}(\xi_2 + \sigma_2) + \frac{S_1^*}{S_2^*}(\xi_1 + \sigma_1)} - \Delta_2, \quad (2.24)$$

where

$$\xi_j = \frac{N_j(\gamma+\mu) - \beta \bar{S}_j^*}{\beta \bar{S}_j^*}, j = 1, 2 \tag{2.25}$$

and

$$\begin{aligned}\Delta_2 = &\frac{\beta_1(1-\sigma_1)I_1^* + \beta_2\sigma_1 I_2^* + \epsilon}{\beta_1(1-\sigma_1)S_1^* + \beta_2(1-\sigma_2)S_2^* - 2(\gamma+\mu)} \frac{\hat{C}_{S_1I_2}^*}{\sqrt{V_1^* V_2^*}} \\ &+ \frac{\beta_2(1-\sigma_2)I_2^* + \beta_1\sigma_2 I_1^* + \epsilon}{\beta_1(1-\sigma_1)S_1^* + \beta_2(1-\sigma_2)S_2^* - 2(\gamma+\mu)} \frac{\hat{C}_{S_2I_1}^*}{\sqrt{V_1^* V_2^*}}.\end{aligned} \tag{2.26}$$

To derive this result we use the moment equation for $\hat{C}_{II}$. By definition, the moment equation for $\mathbb{E}[I_1 I_2]$ is

$$\begin{aligned}\frac{d\mathbb{E}[I_1I_2]}{dt} = \mathbb{E}\Bigg[&\left(\beta S_1\left((1-\sigma_1)\frac{I_1}{N_1} + \sigma_1\frac{I_2}{N_2}\right) + \epsilon S_1\right)(+I_2) + (\gamma+\mu)I_1(-I_2) \\ &+ \left(\beta S_2\left((1-\sigma_2)\frac{I_2}{N_2} + \sigma_2\frac{I_1}{N_1}\right) + \epsilon S_2\right)(+I_1) + (\gamma+\mu)I_2(-I_1)\Bigg],\end{aligned} \tag{2.27}$$

and so

$$\frac{d\hat{C}_{II}}{dt} = \beta_1\sigma_2 S_2 V_1 + \beta_2\sigma_1 S_1 V_2 + \left(\beta_1(1-\sigma_1)S_1 + \beta_2(1-\sigma_2)S_2 - 2(\gamma+\mu)\right)\hat{C}_{II} \tag{2.28}$$

$$+ \left(\beta_1(1-\sigma_1)I_1 + \beta_2\sigma_1 I_2 + \epsilon\right)\hat{C}_{S_1I_2} + \left(\beta_2(1-\sigma_2)I_2 + \beta_1\sigma_2 I_1 + \epsilon\right)\hat{C}_{S_2I_1}. \tag{2.29}$$

At equilibrium $d\hat{C}_{II}/dt = 0$ and if we divide by $\sqrt{V_1^* V_2^*}$, then

$$0 = \sqrt{\frac{V_1^*}{V_2^*}}\beta_1\sigma_2 S_2 + \sqrt{\frac{V_2^*}{V_1^*}}\beta_2\sigma_1 S_1 + \left(\beta_1(1-\sigma_1)S_1 + \beta_2(1-\sigma_2)S_2 - 2(\gamma+\mu)\right)\rho \tag{2.30}$$

$$+ \left(\beta_1(1-\sigma_1)I_1 + \beta_2\sigma_1 I_2 + \epsilon\right)\frac{\hat{C}_{S_1I_2}}{\sqrt{V_1^* V_2^*}} + \left(\beta_2(1-\sigma_2)I_2 + \beta_1\sigma_2 I_1 + \epsilon\right)\frac{\hat{C}_{S_2I_1}}{\sqrt{V_1^* V_2^*}}, \tag{2.31}$$

and hence we have the following approximation for the correlation:

$$\begin{aligned}
\rho = & \sqrt{\frac{V_1^*}{V_2^*}} \frac{-\beta_1\sigma_2 S_2}{\beta_1(1-\sigma_1)S_1 + \beta_2(1-\sigma_2)S_2 - 2(\gamma+\mu)} \\
& + \sqrt{\frac{V_2^*}{V_1^*}} \frac{-\beta_2\sigma_1 S_1}{\beta_1(1-\sigma_1)S_1 + \beta_2(1-\sigma_2)S_2 - 2(\gamma+\mu)} \\
& - \frac{\beta_1(1-\sigma_1)I_1 + \beta_2\sigma_1 I_2 + \epsilon}{\beta_1(1-\sigma_1)S_1 + \beta_2(1-\sigma_2)S_2 - 2(\gamma+\mu)} \frac{\hat{C}_{S_1 I_2}}{\sqrt{V_1^* V_2^*}} \\
& - \frac{\beta_2(1-\sigma_2)I_2 + \beta_1\sigma_2 I_1 + \epsilon}{\beta_1(1-\sigma_1)S_1 + \beta_2(1-\sigma_2)S_2 - 2(\gamma+\mu)} \frac{\hat{C}_{S_2 I_1}}{\sqrt{V_1^* V_2^*}} \qquad (2.32) \\
= & \sqrt{\frac{V_1^*}{V_2^*}} \frac{\beta_1\sigma_2 S_2}{(\gamma+\mu) - \beta_1(1-\sigma_1)S_1 + (\gamma+\mu) - \beta_2(1-\sigma_2)S_2} \\
& + \sqrt{\frac{V_2^*}{V_1^*}} \frac{\beta_2\sigma_1 S_1}{(\gamma+\mu) - \beta_1(1-\sigma_1)S_1 + (\gamma+\mu) - \beta_2(1-\sigma_2)S_2} - \Delta_2 \qquad (2.33) \\
= & \sqrt{\frac{V_1^*}{V_2^*}} \frac{\sigma_2}{\frac{S_1}{S_2}\frac{N_1(\gamma+\mu)-\beta S_1}{\beta S_1} + \frac{S_1}{S_2}\sigma_1 + \frac{N_1}{N_2}\frac{N_2(\gamma+\mu)-\beta S_2}{\beta S_2} + \frac{N_1}{N_2}\sigma_2} \\
& + \sqrt{\frac{V_2^*}{V_1^*}} \frac{\sigma_1}{\frac{S_2}{S_1}\frac{N_2(\gamma+\mu)-\beta S_2}{\beta S_2} + \frac{S_2}{S_1}\sigma_2 + \frac{N_2}{N_1}\frac{N_1(\gamma+\mu)-\beta S_1}{\beta S_1} + \frac{N_2}{N_1}\sigma_1} - \Delta_2 \qquad (2.34) \\
= & \sqrt{\frac{V_2^*}{V_1^*}} \frac{\sigma_1}{\frac{N_2}{N_1}(\xi_1+\sigma_1) + \frac{S_2}{S_1}(\xi_2+\sigma_2)} + \sqrt{\frac{V_1^*}{V_2^*}} \frac{\sigma_2}{\frac{N_1}{N_2}(\xi_2+\sigma_2) + \frac{S_1}{S_2}(\xi_1+\sigma_1)} - \Delta_2. \qquad (2.35)
\end{aligned}$$

2.4 Discussion

A limitation of metapopulation-type models within epidemiological modelling is how to infer the coupling between interacting populations. Sufficiently rich data on relevant interactions is often lacking, especially in developing countries, and it is unclear how such data should translate into a single phenomenological coupling parameter. In light of data on disease incidence being more widely available, we derive an approximation for the correlation, ρ, between the number of infected individuals in two identical populations as a function of the coupling parameter σ, providing a one-to-one mapping between the correlation and the coupling.

The results presented here refine the analysis of (Keeling and Rohani, 2002) and correct an error in the original derivation of ξ. Our numerical results for a measles-like

infection show substantial correlation for all but the weakest coupling. These findings are consistent with similar studies focussing on persistence and spatial synchronisation of measles outbreaks (Lloyd, 2004; Bolker and Grenfell, 1996), despite differences in the characterisation of the basic model. An analytic relationship between the coupling and correlation has been previously derived (Rozhnova et al., 2012) in a more general setting and yields similar numerical results: their relationship is derived through the van Kampen system-size expansion and analysis of the power spectrum. However, we believe that our results provide a significantly simpler relationship between correlations and epidemiological parameters, providing greater intuition and analytical traction. In addition, throughout we compare our analytically tractable results to solution of the moment-based ODEs (given in the Supplementary Information) and to numerical simulation, providing a deeper understanding of the parameter ranges over which the simple results hold and hence the range of applications where the methods are of use. We also differentiate between different modes of failure in our approximation between diseases with low and high basic reproductive ratios.

Our work also offers an alternative parametrisation of ξ (Equation (2.11)) that depends only on the epidemiological parameters and holds in the large-population limit; however, our numerical analysis shows that this parametrisation also leads to a good qualitative approximation in populations of size $N = 10^3$. This parametrisation is preferable to the original as it provides intuition and insight into how the epidemic parameters affect the correlation. In addition, it removes the need to estimate the number of susceptible individuals in the population at endemic equilibrium, either from data or through simulation. This is particularly useful in smaller populations where we find that the MVN moment closure approximation fails numerically and the solution to the ODEs 'blows up'. This type of failure is well-documented in the literature and typically attributed to large negative covariances, frequent global extinctions or when the distribution of states is bimodal (Keeling, 2000a; Lloyd, 2004; Krishnarajah et al., 2005; Keeling, 2000b; Nasell, 1999). In the limit as $N \to \infty$, (Nasell, 1999) shows that the distribution of states conditioned on non-extinction is approximately normal when R_0 is greater than 1; however, this does not explain why the MVN moment closure approximation sometimes appears to hold for smaller populations, such as in our own analysis and the wider literature (Isham, 1995).

Our model is sufficiently general that it can describe multiple forms of heterogeneity in the population including spatial, age and risk heterogeneity; however, a limitation of the model it that the underlying SIR model is too simple to describe the full dynamics

of many diseases. For example, as noted previously a full model of measles dynamics should include both seasonality and age structure. These limitations should be addressed before using our results to infer the between-population coupling parameters. We have also shown that adding complexity reduces the analytical tractability of the model, such as with populations of unequal size; in the most general case, analysis of the model may require a computational, rather than analytical, approach. Finally, whilst data on disease incidence in each of the populations is more widely available than mobility data, our results are still limited by the availability and quality of such data. In particular, our results will be affected by under-reporting of infections.

Our results provide a method by which the coupling can be estimated from the correlation between the number of infected individuals in two populations using data on disease incidence. Crucially, this allows us to estimate the coupling between populations even in the absence of data on human mobility, thus circumventing one of the main limitations of metapopulation models. At present our model considers the mathematically tractable case of two identical populations at endemic equilibrium. Future research should aim to address the limitations outlined above by improving the underlying epidemic model, for example by incorporating seasonality or age structure. The current model can easily be extended to multiple identical interacting populations when the underlying graph is a symmetric graph, such as the complete graph or k-regular infinite tree graph. This holds since any adjacent populations will have identical neighbourhoods. These extensions will significantly improve the realism of the model and validate the use of the results in the inference of between-population coupling parameters.

2.5 Conclusions

A limitation of metapopulation models is how to infer the coupling between interacting populations. In this chapter, we relate the correlation between the number of infected individuals in two identical populations as a function of the coupling, providing a one-to-one mapping between the correlation and the coupling. Combined with data on disease incidence in each of the populations, this result provides a method by which the between-population coupling can be estimated, even in the absence of information on the population mobility.

Chapter 3

Estimating the between-subpopulation coupling from the correlation

3.1 Introduction

An ongoing challenge in metapopulation modelling of infectious diseases is how to infer the coupling between subpopulations. The individual-level behaviour that determines the interactions between groups is highly complex and is dependent on socio-demographic factors (Mossong et al., 2008; González et al., 2008; Horby et al., 2011; Danon et al., 2013; Read et al., 2014; Stopczynski et al., 2014; Wesolowski et al., 2015; Kiti et al., 2016; Klepac et al., 2018) and good data on relevant interactions are not always readily available. Moreover, even with access to good data on relevant interactions, it is unclear how this should translate into a transmission parameter.

On the other hand, long-term data on disease incidence is often more widely available (Olsen and Schaffer, 1990; Grenfell and Harwood, 1997). In Chapter 2, we derived an approximation for the correlation, ρ, between the prevalence of infected individuals in two identical subpopulations as a function of the coupling parameter σ, providing a one-to-one mapping between the correlation and the coupling. Therefore, given we can observe the correlation between prevalence of infection in two subpopulations, we proposed that we can use this approximation to estimate the coupling between the subpopulations.

In this section we aim to explore whether this method to estimate the coupling is

feasible. Under different limitations to the observation process, we compare the true coupling to our estimate of the coupling using the approximation from Chapter 2. First we consider the best-case scenario, where there are no limitations to the observation process. Then we explore the unilateral effect of three limitations: a shorter observation period, less frequent observations, and only observing incidence of new infections or recoveries. Finally, we consider the effect of all three limitations combined. Throughout this chapter we use parameter values for a measles-like endemic disease in the UK.

3.2 Simulation of stochastic processes

We use the stochastic model of endemic infection for two coupled subpopulations, as defined in Chapter 2, Section 2.2.2; the transition rates for this process are summarised in Table 2.1. We simulate the Markov process using the τ-leaping algorithm ($\tau = 1$) (defined in Chapter 1, Section 1.2.4). For each realisation of the process, we calculate the correlation between the infection prevalence in the two subpopulations as the time-weighted Pearson correlation coefficient. For each value of the coupling $\sigma \in \{10^{-4}, 10^{-3.8}, \ldots, 10^{-0.2}, 1\}$ we generate 1000 realisations of the process. We use epidemiological parameters representing a measles-like endemic disease in the UK: $N = 10^5$, $R_0 = 17$, $\gamma^{-1} = 13$ days, $\mu = 5.5 \times 10^{-5}$ days^{-1} and $\epsilon = 5.5 \times 10^{-5}$ days^{-1}.

The observation process is the part of the Markov process that we are able to observe, and is defined by the length of the observation period T, the frequency of observations, and the state of the process that we observe. As a base case, we assume that the length of the observation period is $T = 200$ years, and that we make daily observations of the infection prevalence (that is, of I_1 and I_2).

3.3 Estimating the coupling with perfect data

First we consider whether it is feasible to estimate the coupling from the correlation when there are no limitations to the observation process. This represents a best-case scenario where daily infection prevalence in the two subpopulations is over a long period of time ($T = 200$ years).

3.3.1 Estimating the true coupling

In Chapter 2 we derived an approximation for the correlation, ρ, between infection prevalence in two identical subpopulations in terms of the coupling, σ, between them:

$$\rho \approx \frac{\sigma}{\xi' + \sigma},$$

where $\xi' = \epsilon/(\mu(R_0 - 1))$. Therefore, given we observe correlation $\hat{\rho}$, we can therefore estimate the true coupling by $\hat{\sigma}$:

$$\hat{\sigma} = \min\left(\frac{\xi'\hat{\rho}}{(1-\hat{\rho})}, 1\right). \tag{3.1}$$

where we take the minimum so that $\hat{\sigma} \in [0, 1]$. For each realisation of the stochastic simulations we use this to calculate the estimated coupling based on the observed correlation.

3.3.2 Comparing true and estimated coupling

To evaluate this method for estimating the coupling we compare the estimated coupling $\hat{\sigma}$ and the true coupling σ. We also calculate the absolute error of our estimate as $\hat{\sigma} - \sigma$.

We can estimate the true coupling reasonably well using the method, although in general we underestimate the true coupling σ (Figure 3.1). This is because our approximation $\sigma/(\xi' + \sigma)$ overestimates the correlation for a given coupling. For low coupling values ($\sigma < 0.01$) we are not able to consistently estimate the true coupling correctly: there is considerable variation in the estimated coupling (as shown by the width of the error bars), although the absolute error is small. On the other hand, for high coupling values we consistently underestimate the true coupling: the variation in the estimates is low, but the absolute error is large. However, we should note that the largest coupling values ($\sigma > 0.5$) are somewhat unrealistic, since they represent a metapopulation where there is more interaction between subpopulations than within subpopulations.

In summary, we are able to estimate the coupling from the correlation between infection prevalence when there are no limitations to the observation process. Although the absolute error in the estimated coupling is too high for some values of the true coupling, these high coupling values correspond to metapopulations that are unrealistic in a real-world setting.

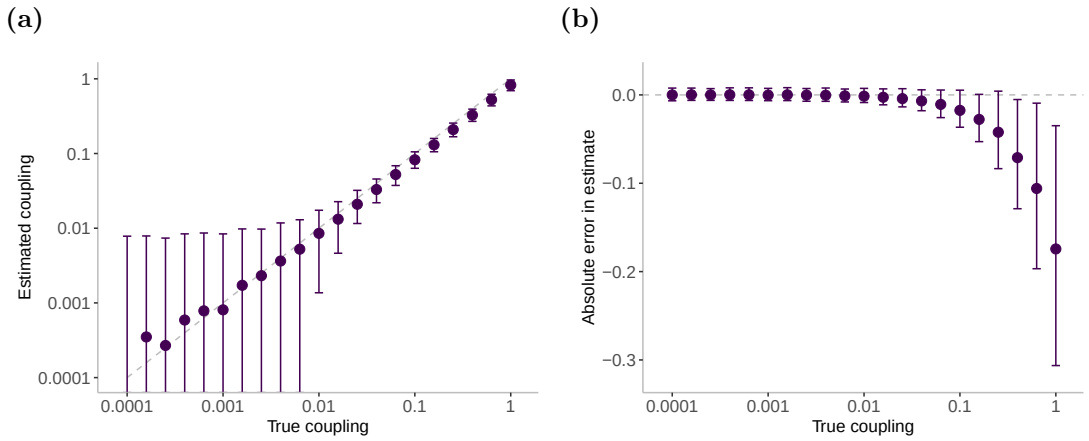


Figure 3.1. Evaluating our estimate of the coupling. **(a)** Comparing the true coupling and the estimated coupling, $\hat{\sigma} = \xi'\hat{\rho}/(1-\hat{\rho})$. Dashed line indicated $y = x$, for reference. **(b)** The absolute error in the estimated coupling. Dashed line shows zero error, for reference. In both figures points represent the mean taken over 1000 realisations and error bars show 2.5th and 97.5th percentiles.

3.4 Estimating the coupling with limited data

In a real-world setting there will likely be some limitations to our observations of the infectious disease process, which may affect the observed correlation and our subsequent estimate of the coupling. We consider the effect of three limitations to the observation process. First, we unilaterally consider the effect of a shorter observation period, less frequent observations, and observing incidence (rather than prevalence) data. Then we combine these limitations, which represents a more realistic observation process.

3.4.1 Shorter observation period

In the best-case scenario we observed the data over a 200-year period. We now consider the effect of a shorter observation period by limiting our observation period to intervals of 1 year up to 10 years, and then multiples of 5 years up to 50 years. As in the best-case scenario, we make daily observations and observe infection prevalence in each subpopulation.

Effect on the correlation

First we calculate the observed correlation for the limited process and compare it to the true correlation (that is, the correlation when the process is observed for the full 200 years) (Figure 3.2). For very low values of the coupling ($\sigma = 0.001$) and a short observation period ($T < 10$ years) the observed correlation is an overestimate of the true correlation; on the other hand, for larger values of the coupling ($\sigma = 0.1$) and the same short observation period, the observed correlation is an underestimate of the true correlation. For all values of σ the variability in the observed correlation is large for short observation periods, but this variability decreases as the length of the observation period increases.

Effect on the estimated coupling

We estimate the coupling when the observation period is $T = 10$, 30 and 50 years, and compare the estimates to the true coupling (Figure 3.3). For all but the lowest true coupling values, the mean estimated coupling is comparable to the true coupling. For very low coupling values ($\sigma \leq 0.001$) and a short observation period ($T = 10$ years), we overestimate the true coupling; as we showed in Figure 3.2, this is because we tend to overestimate the correlation under these conditions. As the length of the observation period increases, these estimates improve. Similarly, as the length of the observation period increases then the variability in the estimated coupling decreases; this is because the variability in the observed correlation decreases as the length of the observation period increases. The mean absolute error in the estimated coupling for the limited process is comparable to the mean absolute error in the full process; however, the variability is much higher for a short observation period, again as a result of the variability in the observed correlation.

3.4.2 Lower frequency observations

In the best-case scenario we observed the daily infection prevalence. We consider the effect of less frequent observations by instead observing infection prevalence every 7, 30 and 90 days (weekly, approximately monthly, and approximately every three months). We then calculate the correlation between infection prevalence in the two subpopulations using these 'thinned' observations. As in the best-case scenario, we observe the process for $T = 200$ years and observe the infection prevalence in each subpopulation.

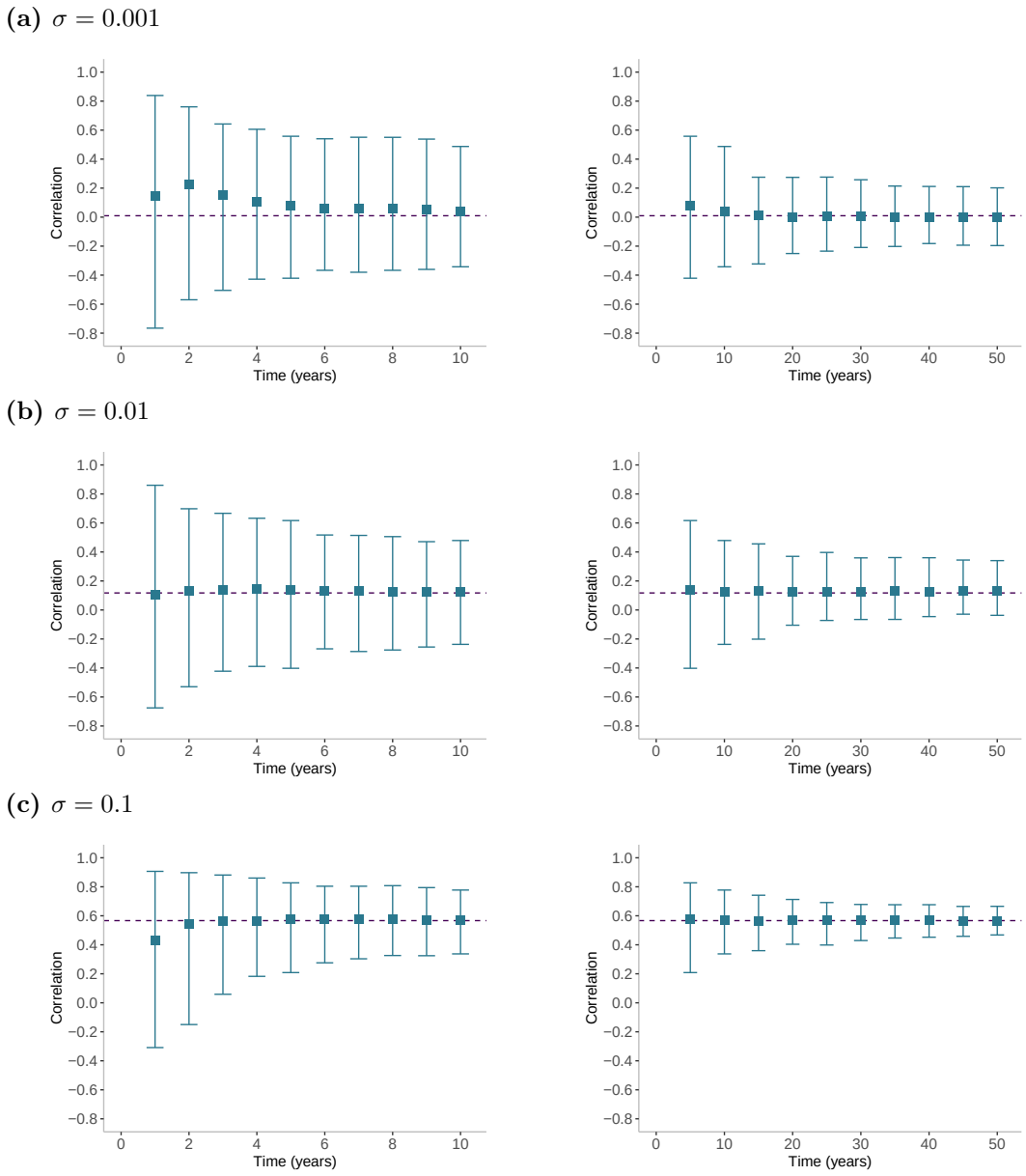


Figure 3.2. Effect of a shorter observation period on the correlation between infection prevalence in two subpopulations, for $\sigma \in \{0.001, 0.01, 0.1\}$. The observation period, T, is multiples of 1 year up to 10 years (left), and then multiples of 5 years up to 50 years (right). Points show the mean correlation taken over 1000 realisation, and error bars show 2.5th and 97.5th percentiles. Dashed line shows the mean correlation when the process observed for 200 years, for reference.

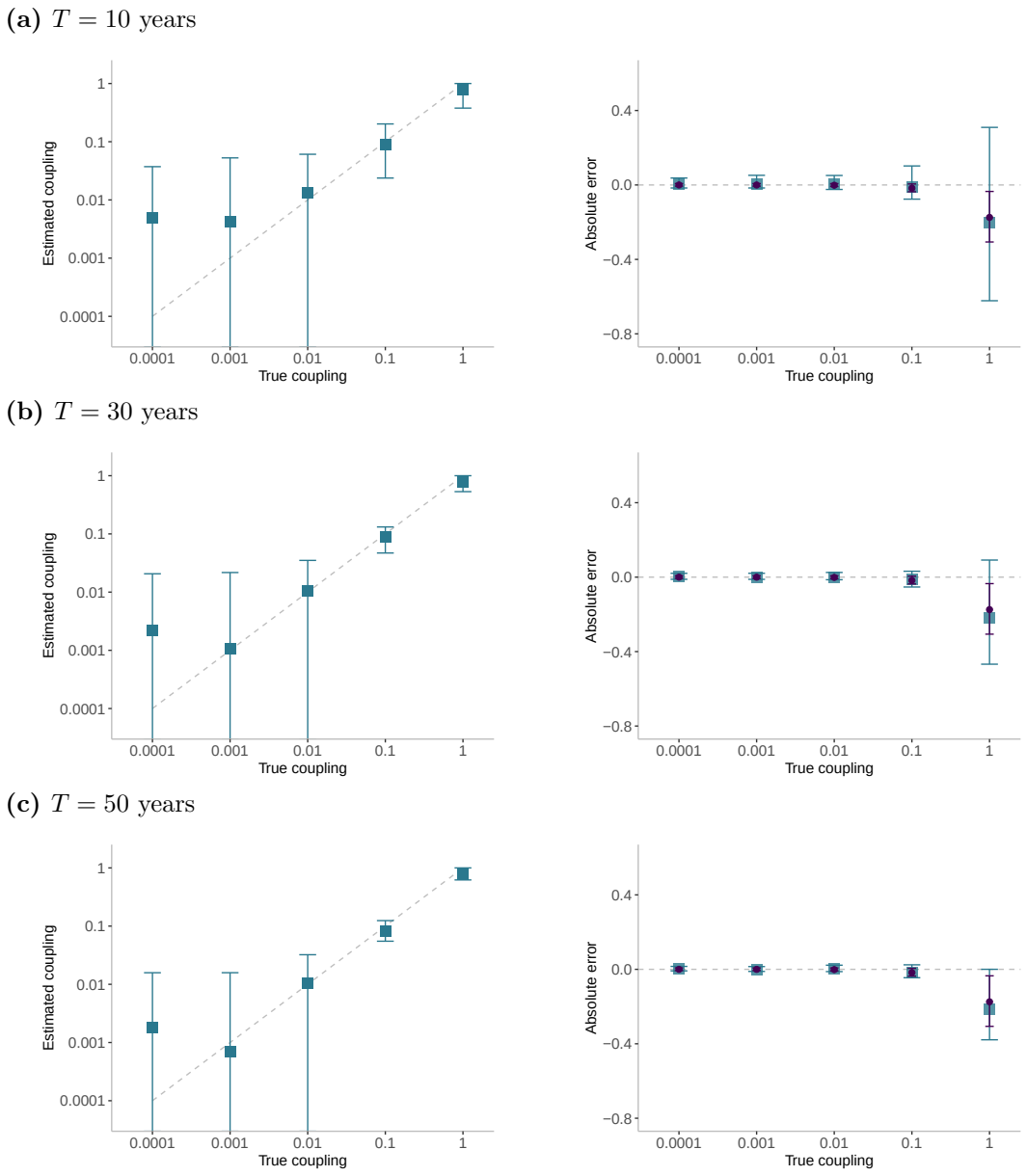


Figure 3.3. Evaluating the effect of a shorter observation period, T, on our estimate of the coupling, for $T = 10, 30, 50$ years. Left panel compares the true coupling and the estimated coupling, $\hat{\sigma} = \xi'\hat{\rho}/(1-\hat{\rho})$; dashed line shows $y = x$, for reference. Right panel shows the absolute error in the estimated coupling for the process with $T < 200$ years (square points) and for $T = 200$ years (circle points, for reference); dashed line shows zero error, for reference. In all figures, points represent the mean taken over 1000 realisations and error bars show 2.5th and 97.5th percentiles.

Effect on the correlation

We calculate the observed correlation for the limited processes and compare to the true correlation (that is, the process where observations are made daily). Lower frequency observations have very little effect on the observed correlation for all values of the coupling σ (Figure 3.4). Both the mean observed correlation and the variability in the observed correlation is almost indistinguishable from the true correlation.

Effect on the estimated coupling

Since lower frequency observations have no noticeable effect on the observed correlation, then this limitation also has little effect on the estimated coupling (Figure 3.5). The absolute error in the estimated coupling for the limited processes is comparable to the absolute error for the full process. Even when observations are only made every 90 days, there is only a small increases in the variability of the absolute error in comparison to the absolute error for the full process.

3.4.3 Incidence data

In the best-case scenario we assumed that we observed the infection prevalence. However, in practice we are more likely to be able to observe the incidence of new cases or recoveries. In this section we consider the effect of observing the recovery incidence (that is, the number of individuals moving from the I to R class) in the two subpopulations. As in the best-case scenario, we make daily observations and observe the process for $T = 200$ years.

Effect on the correlation

The correlation between recovery incidence is much lower than the correlation between infection prevalence (Figure 3.6). However, we can show that the correlation between recovery incidence, which we will denote by ρ_X, is related to the correlation between infection prevalence, ρ_I, in the following way:

$$\rho_I = \frac{Var(X)}{Var(X) - \mathbb{E}[X]}\rho_X,$$

where $\mathbb{E}[X]$ and $Var(X)$ are the mean and variance in recovery incidence, respectively.

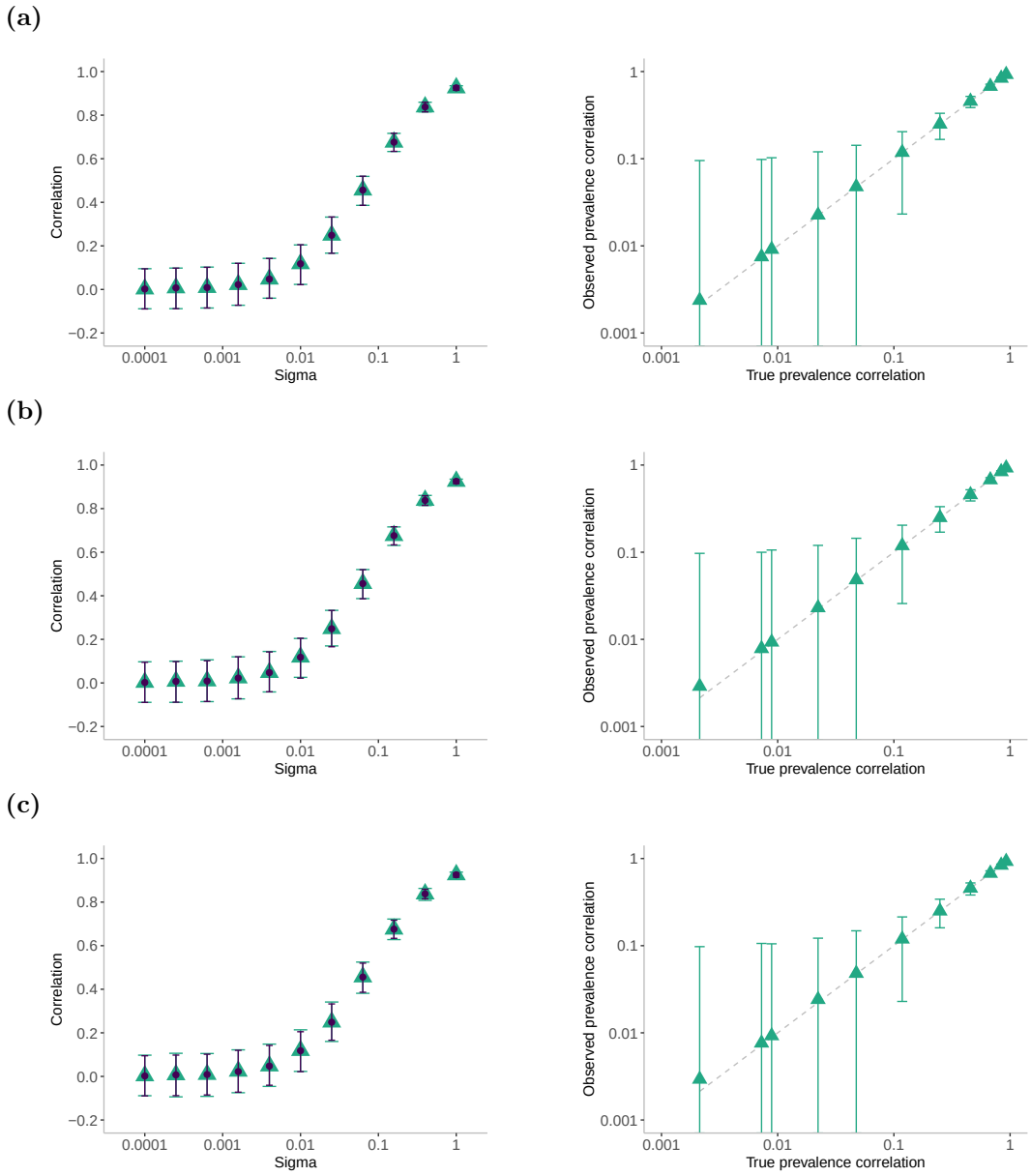


Figure 3.4. Effect of less frequent observations on the correlation between infection prevalence in two subpopulations. We observe the process every ((**a**)-(**c**)) 7, 30 and 90 days. Left panel shows the prevalence correlation for the thinned process (triangle points) and unthinned process (circle points, for reference) for coupling $\sigma \in [0, 1]$. Right panel directly compares the true correlation and the observed correlation; dashed line shows $y = x$, for reference. In all figures, points represent the mean taken over 1000 realisations and error bars show 2.5th and 97.5th percentiles.

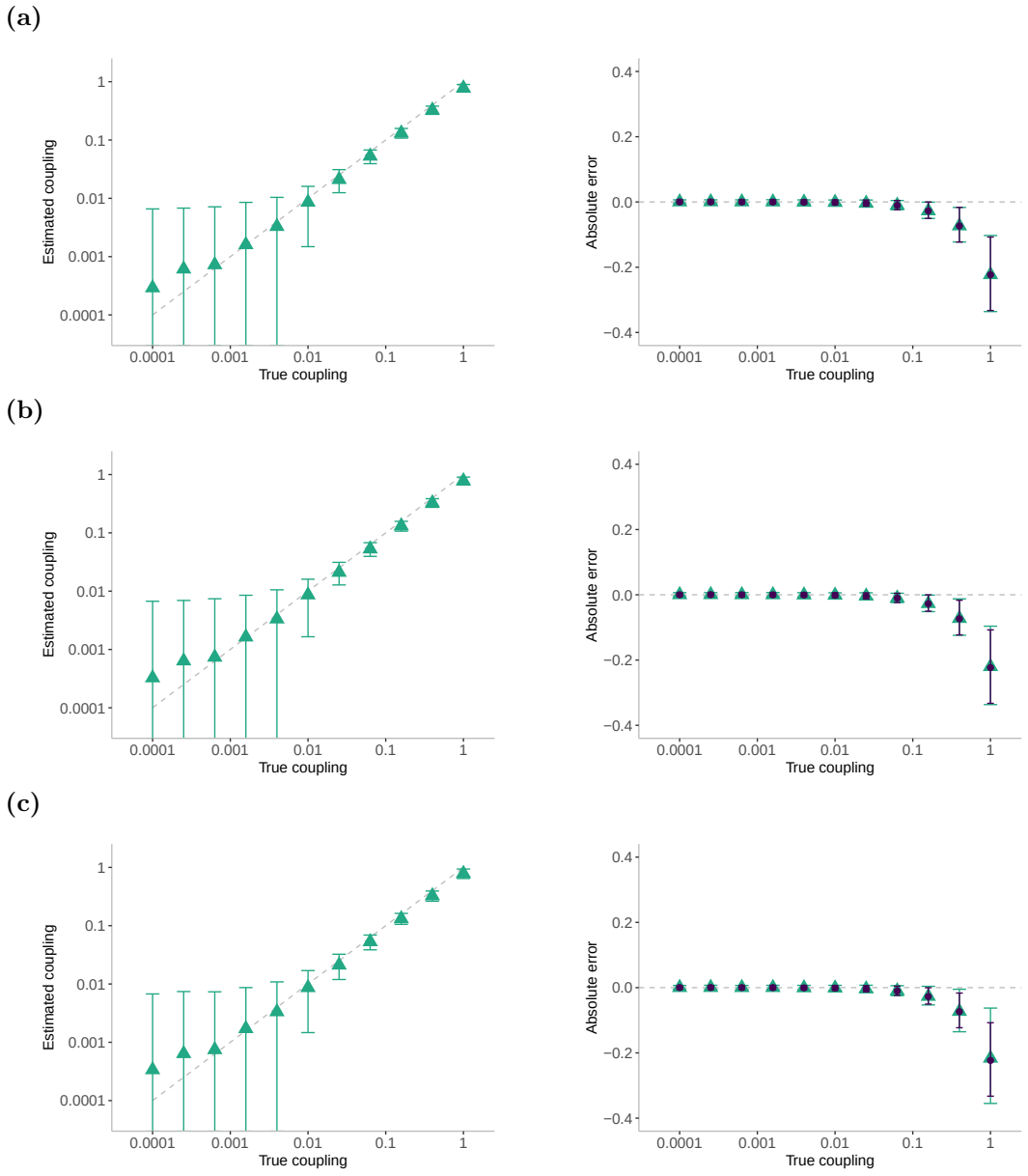


Figure 3.5. Evaluating the effect of lower frequency observations on our estimate of the coupling, where observations are made every (**(a)-(c)**) 7, 30 and 90 days. Left panel compares the true coupling and the estimated coupling, $\hat{\sigma} = \xi\hat{\rho}/(1-\hat{\rho})$; dashed line shows $y = x$, for reference. Right panel shows the absolute error in the estimated coupling for the thinned process (triangle points) and unthinned process (circle points, for reference); dashed line shows zero error, for reference. In all figures points represent the mean taken over 1000 realisations and error bars show 2.5th and 97.5th percentiles.

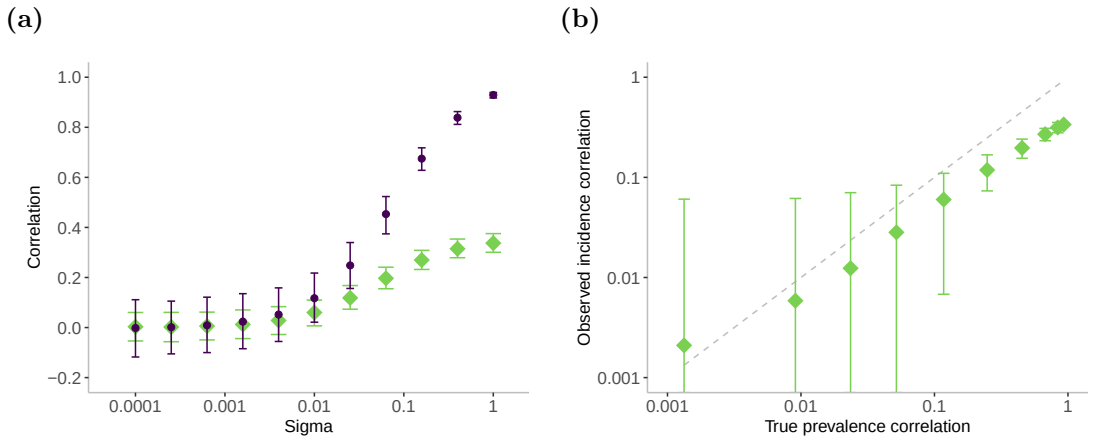


Figure 3.6. Effect of observing recovery incidence on the correlation. **(a)** Comparing the correlation between recovery incidence (diamond points) and the correlation between infection prevalence (circle points, for reference) for coupling $\sigma \in [0, 1]$. **(b)** Directly comparing correlation between infection prevalence and correlation between recovery incidence; dashed line shows $y = x$, for reference. In both figures points represent the mean correlation taken over 1000 realisations; error bars represent 2.5th and 97.5th percentiles.

To show this result, let X_j be the daily recovery incidence in subpopulation $j = 1, 2$. We can write down the distribution of X_j: since individuals recover at the points of a Poisson process with rate γI_j, then $X_j \sim \text{Poisson}(\gamma I_j)$. As the two subpopulations are identical, the expected behaviour of X_1 and X_2 is the same, that is, $\mathbb{E}[X_1] = \mathbb{E}[X_2]$ and $Var(X_1) = Var(X_2)$. By properties of the conditional expectation, the mean of $X_j, j = 1, 2$ is given by

$$\begin{aligned} \mathbb{E}[X] = \mathbb{E}[X_1] &= \mathbb{E}\Big[\mathbb{E}[X_1|I_1]\Big] \\ &= \gamma \mathbb{E}[I_1] \\ &= \gamma \bar{I}, \end{aligned} \tag{3.2}$$

and by the law of total variance, the variance of $X_j, j = 1, 2$ is given by

$$\begin{aligned} Var(X) = Var(X_1) &= Var\Big(\mathbb{E}[X_1|I_1]\Big) + \mathbb{E}\Big[Var(X_1|I_1)\Big] \\ &= Var(\gamma I_1) + \mathbb{E}[\gamma I_1] \\ &= \gamma^2 Var(I) + \gamma \mathbb{E}[I] \\ &= \gamma^2 C_{II} + \gamma \bar{I}. \end{aligned} \tag{3.3}$$

The correlation between the recovery incidence in the two subpopulations is defined as

$$\rho_X = \frac{cov(X_1, X_2)}{\sqrt{Var(X_1)Var(X_2)}} = \frac{cov(X_1, X_2)}{Var(X)}.$$

Using properties of the conditional expectation and Equation (3.2), we can show that $cov(X_1, X_2) = \gamma^2 cov(I_1, I_2) = \gamma^2 \hat{C}_{II}$:

$$\begin{aligned} cov(X_1, X_2) &= \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] \\ &= \mathbb{E}\Big[\mathbb{E}[X_1 X_2|I_1, I_2]\Big] - \gamma^2 \bar{I}^2 \\ &= \mathbb{E}\Big[\mathbb{E}[X_1|I_1, I_2]\mathbb{E}[X_2|I_1, I_2]\Big] - \gamma^2 \bar{I}^2 \\ &= \mathbb{E}[\gamma I_1 \gamma I_2] - \gamma^2 \bar{I}^2 \\ &= \gamma^2 \Big(\mathbb{E}[I_1 I_2] - \bar{I}^2\Big) \\ &= \gamma^2 \hat{C}_{II}. \end{aligned} \tag{3.4}$$

Therefore, combining Equation (3.3) and Equation (3.4), the correlation ρ_X is written as:

$$\begin{aligned} \rho_X &= \frac{\gamma^2 \hat{C}_{II}}{\gamma^2 C_{II} + \gamma \bar{I}} \\ &= \frac{\hat{C}_{II}}{C_{II} + \bar{I}/\gamma} \\ &= \frac{C_{II}}{C_{II} + \bar{I}/\gamma} \frac{\hat{C}_{II}}{C_{II}} \\ &= \frac{C_{II}}{C_{II} + \bar{I}/\gamma} \rho_I. \end{aligned}$$

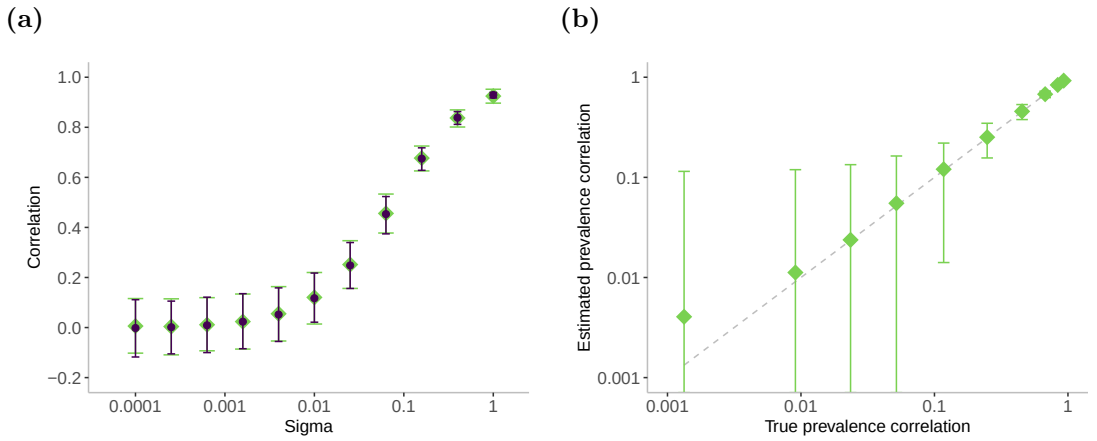


Figure 3.7. Comparing the true prevalence correlation and the estimated prevalence correlation, estimated using incidence observations. **(a)** Comparing the true prevalence correlation (circle points) and the estimated prevalence correlation (diamond points) for coupling $\sigma \in [0, 1]$. **(b)** Directly comparing the true prevalence correlation and the estimated prevalence correlation; dashed line shows $y = x$, for reference. In both figures points represent the mean taken over 1000 realisations; error bars represent 2.5th and 97.5th percentiles.

Moreover, since $Var(X) = \gamma^2 C_{II} + \gamma \bar{I}$ and $\mathbb{E}[X] = \gamma \bar{I}$, then we have

$$\begin{aligned} \rho_X &= \frac{Var(X) - \mathbb{E}[X]}{Var(X)} \rho_I \\ \iff \rho_I &= \frac{Var(X)}{Var(X) - \mathbb{E}[X]} \rho_X. \end{aligned} \tag{3.5}$$

Therefore, given that we observe the incidence time series $X_j, j = 1, 2$, then we can directly estimate the correlation between infection prevalence using Equation (3.5). We compare the true prevalence correlation to the prevalence correlation estimated from the observations of recovery incidence (Figure 3.7), from which we observe that there is very little difference between the two correlation values.

Effect on the estimated coupling

Since we can obtain very accurate estimates of the prevalence correlation by using observations of recovery incidence, then this limitation also has little effect on the estimated coupling (Figure 3.8). The mean absolute error in the estimated coupling from the inci-

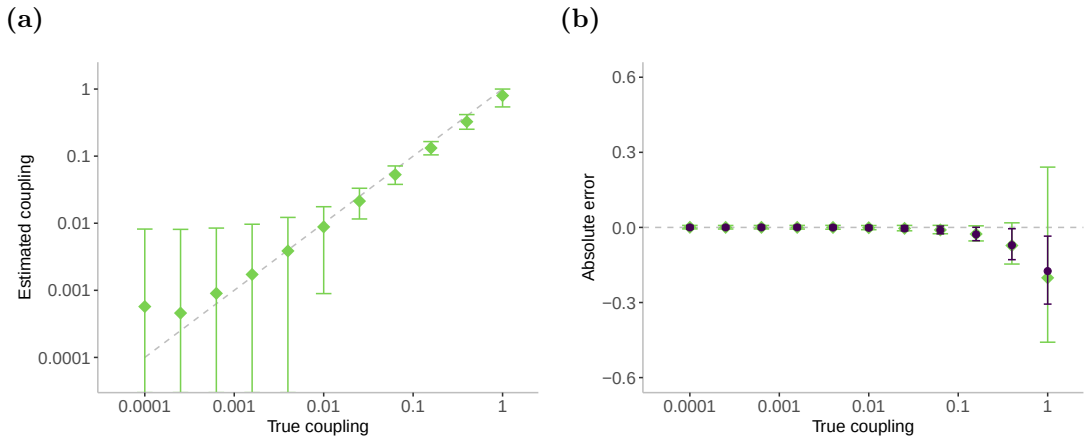


Figure 3.8. Evaluating the effect of observing incidence of infection on our estimate of the coupling. **(a)** Comparing the true coupling and the estimated coupling, $\hat{\sigma} = \xi\hat{\rho}/(1-\hat{\rho})$; dashed line shows $y = x$, for reference. **(b)** Showing the absolute error in the estimated coupling incidence observations (diamond points) and prevalence observations (circle points, for reference); dashed line shows zero error, for reference. In all figures points represent the mean taken over 1000 realisations and error bars show 2.5th and 97.5th percentiles.

dence observations are comparable to the mean absolute error for the estimated coupling from the prevalence observations; the only difference is that for the highest coupling values ($\sigma = 1$), the variability in the absolute error is larger for incidence observations compared to prevalence observations.

3.4.4 Realistic time series data

We combine these above limitations to represent a more realistic observation process. We assume that the observation period is $T = 25$ years and that we observe recovery incidence. Observations are thinned so that we observe the aggregated incidence every 7 days, that is, every 7 days we observe the sum of the daily incidence counts.

Effect on the correlation

We calculate the observed correlation between recovery incidence for the realistic process, and then using Equation (3.5) we estimate the correlation between infection prevalence in the two subpopulations (Figure 3.9). For a short observation period ($T = 25$ years) the

estimated correlation is consistently higher than the true correlation, and is significantly higher for the very high coupling values ($\sigma \approx 0.4$). For a long observation period ($T = 200$ years) the estimated correlation improves for low coupling values, but still overestimates the true correlation for high coupling values. This is particularly surprising since neither weekly observations nor observing recovery incidence had this effect on the correlation when considered individually. As we might expect given the results for a shorter observation period (Section 3.4.1), the variability in the estimated prevalence correlation appears to be determined by the length of the observation period. When $T = 25$ years, the variability in the estimated prevalence correlation is high, and when $T = 200$ then the variability in the estimated prevalence correlation is comparable to when there are no limitations on the observation process.

Effect on the estimated coupling

In general, our method to estimate the coupling performs well even for an observation process with multiple limitations. However, at high levels of coupling, the combined effect of the limitations on the estimated coupling is striking, and we significantly overestimate the true coupling (Figure 3.10). When $\sigma \approx 0.4$, the mean estimated coupling is 0.962 for $T = 25$ years, and 0.995 for $T = 200$ years (that is, the mean absolute error is approximately 0.6). We also overestimate the coupling for low coupling values and a short observation period ($T = 25$ years), although in this case the absolute error is very small and the estimate improves when the length of the observation period increases to $T = 200$ years.

3.5 Discussion

An ongoing challenge in metapopulation modelling of infectious diseases is how to infer the coupling between subpopulations. In this chapter we consider the feasibility of estimating the coupling from the correlation between the infection prevalence in two interacting subpopulations, using the analytic relationship derived in Chapter 2. In particular, we consider how several realistic limitations on the observation process affect the correlation and subsequent estimates of coupling, namely a shorter observation period, less frequent observations, and observing recovery incidence than infection prevalence.

The work complements the results derived in Chapter 2. We use the simple sigmoidal relationship between the coupling and the correlation to get an estimate of the coupling

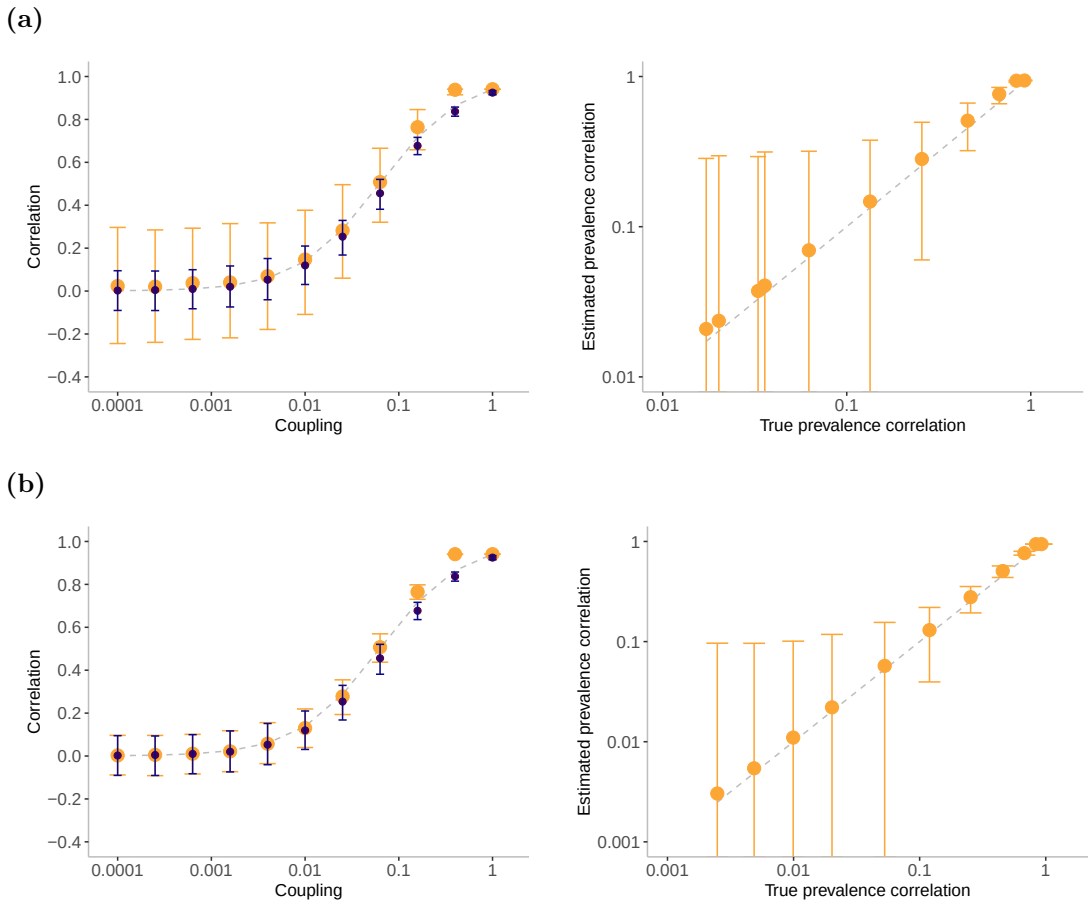


Figure 3.9. Effect of a realistic observation process on the correlation, where the length of the observation period is **(a)** $T = 25$ years, and **(b)** $T = 200$ years. Left panel compares the prevalence correlation for the realistic observation process (big circle points) and the observation process with no limitations (small circle points, for reference) for coupling $\sigma \in [0, 1]$. Right panel directly compares the true prevalence correlation and the prevalence correlation estimated from the realistic observation process; grey dashed line shows $y = x$, for reference. In all figures points represent the mean correlation taken over 1000 realisations; error bars represent 2.5th and 97.5th percentiles.

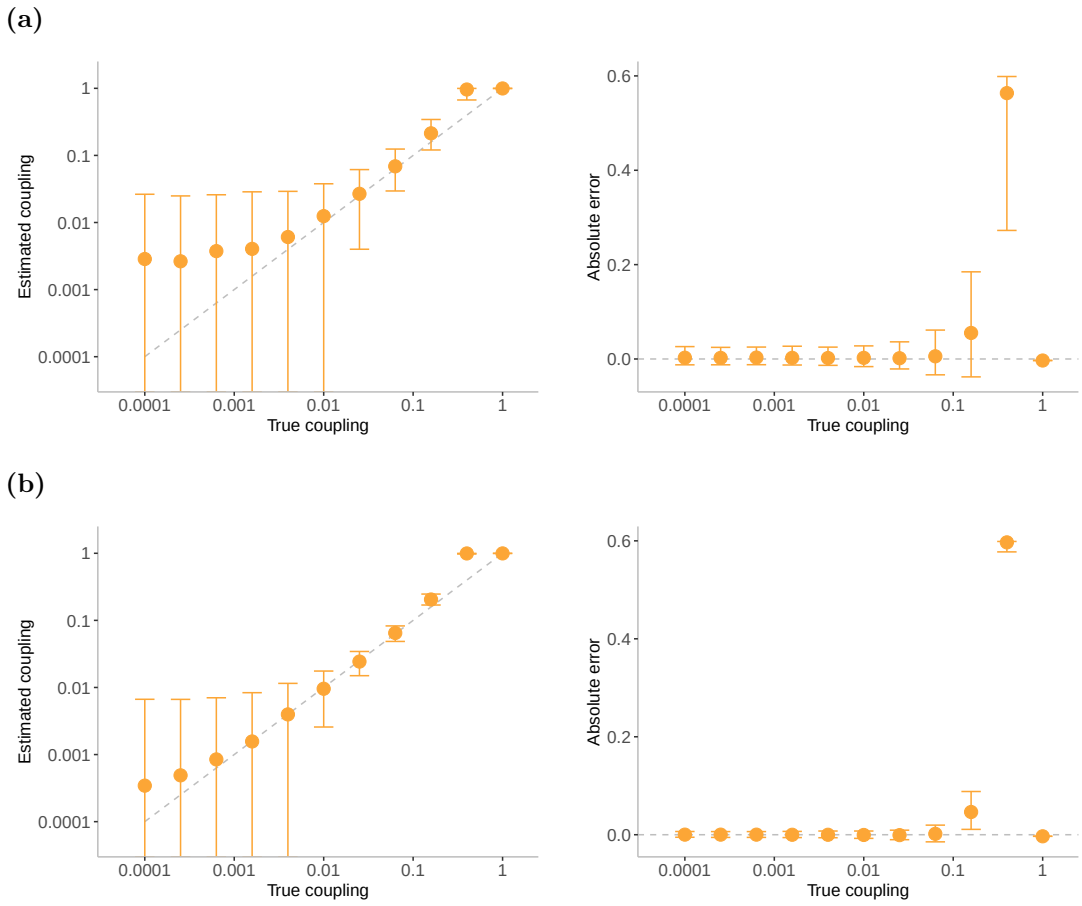


Figure 3.10. Evaluating the effect of a realistic observation process on our estimate of the coupling, where the length of the observation process is **(a)** $T = 25$ years, and **(b)** $T = 200$ years. Left panel compares the true coupling and the estimated coupling, $\hat{\sigma} = \xi\hat{\rho}/(1-\hat{\rho})$; dashed line shows $y = x$, for reference. Right panel shows the absolute error in the estimated coupling for the realistic observation process (big circle points) and the observation process with no limitations (circle points, for reference); dashed line shows zero error, for reference. In all figures points represent the mean taken over 1000 realisations and error bars show 2.5th and 97.5th percentiles.

given we observe the infection prevalence in the two subpopulations. Given no limitations to the observation process, we are able to estimate the true coupling with reasonable accuracy for all but the very largest coupling values, although we emphasise that high coupling values do not represent what we would typically think of as metapopulation dynamics. This supports the suggestion in Chapter 2 that this relationship could be used estimate the coupling between subpopulations, even in the absence of contact or mobility data. The accuracy of the estimated coupling is limited by the accuracy of the approximation $\rho \approx \sigma/(\xi'+\sigma)$. Clearly, the absolute error in the estimated coupling could be reduced by using the exact relationship between the coupling and the correlation: $\rho = \sigma/(\xi+\sigma) - \Delta$. However, since Δ is a function of the covariance $cov(S_1, I_2)$, which we are unlikely to observe, then this is not a realistic solution.

By considering the effect of limitations on the observation process we are able to show whether this approach would be feasible using real-world data. We show that both lower frequency observations (up to every 90 days) and observing recovery incidence have little effect on the estimated coupling when considered separately. Conveniently, we are able to estimate the correlation between infection prevalence even if we only observe the daily recovery incidence. However, when both limitations are applied together then our method for estimating the coupling is less effective: at high coupling values we significantly overestimate the true coupling, even though this effect is not seen by either limitation in isolation. Understanding this behaviour requires further analysis, but we hypothesise that by aggregating daily incidence data, fluctuations in incidence are 'smoothed out' and so $Var(X)/(Var(X) - \mathbb{E}[X])$ is too big.

Further analyses would strengthen the results shown here and thus provide further support for using this method in a real-world setting. We can consider other limitations to the observation process, such as the effect of underreporting or unobserved cases, or observing incidence of infection. We can easily include underreporting by using a binomial thinning on observation of the recovery incidence, but we expect that observing infection incidence will be less simple to resolve than observing recovery incidence, since the rate of the Poisson process for infection events is more complex. This work would also benefit from further theoretical results on the interaction between different limitations, such as observing aggregated recovery incidence. In addition, we might also want to consider how these results are affected by more general metapopulation network configurations.

This chapter strongly supports our assertion in Chapter 2 that we can use our approximation for the correlation to estimate the coupling between subpopulations in a

metapopulation network. Further research into the effect of limitations to the observation process would improve our understanding of the conditions under which the method can be used, and so addresses an ongoing challenge in metapopulation modelling of infectious diseases, namely the inference of coupling between subpopulations.

Chapter 4

Correlations between stochastic endemic infection symmetric metapopulation networks

4.1 Introduction

In this chapter we extend the results derived in Chapter 2 to more general metapopulation networks. Using a multivariate normal approximation, we derive an approximation for the correlation between the infection prevalence in two subpopulations in symmetric metapopulation networks, as a function of the coupling between them. We derive results for subpopulations arranged on the complete network, the k-regular tree network and the star network. We also numerically validate our model by comparing our analytic approximations to stochastic simulations. These results also provide initial insights into the effect of metapopulation network structure on network correlations.

4.2 A stochastic endemic infection model for interacting populations on a general graph

We extend the stochastic endemic infection model for two interacting subpopulations described in Chapter 2, Section 2.2.2, to P subpopulations in a general metapopulation network. We assume throughout the chapter that the population sizes are equal for

Population	Event	Transition	Rate
$j = 1, 2, \ldots, P$	Infection	$s_j \to s_j - 1, i_j \to i_j + 1$	$\beta s_j \sum_l \sigma_{jl} i_l / N + \epsilon s_j$
	Recovery	$i_j \to i_j - 1, r_j \to r_j + 1$	γi_j
	Death of infected	$s_j \to s_j + 1, i_1 \to i_j - 1$	μi_j
	Death of recovered	$s_j \to s_j + 1, r_j \to r_j - 1$	$\mu(N - s_j - i_j)$

Table 4.1. A summary of the transition rates of the $2P$-dimensional Markov chain endemic infection model $\{(S_j(t), I_j(t))_{j=1}^P : t \geq 0\}$ from state $(s_1, i_1, s_2, i_2, \ldots, s_P, i_P)$ with birth/death rate $\mu > 0$, contact rate $\beta > 0$, external import rate $\epsilon > 0$, recovery rate $\gamma > 0$ and coupling matrix $\mathbf{\Sigma}$.

mathematical tractability. Each population exhibits the same population dynamics as the simple model of endemic infection (Chapter 2, Section 2.2.1), plus pairwise interaction between the populations: we assume that in population i, a proportion $\sigma_{ij} \in [0, 1]$ of an individual's contacts are with individuals in population j. We insist that $\sum_j \sigma_{ij} = 1$ and so $\sigma_{ii} = 1 - \sum_{j \neq i} \sigma_{ij}$. The matrix $\mathbf{\Sigma} = (\sigma_{ij})$ therefore describes the interaction or 'coupling' between all possible pairs of populations, and the force of infection in each subpopulation depends on the number of infected individuals in all other subpopulations. Changing $\mathbf{\Sigma}$ does not change the basic reproductive ratio, but instead determines the distribution of secondary cases between the P subpopulations.

We let $S_i(t), I_i(t), R_i(t) \in \{0, 1, 2, \ldots\}$ denote the number of susceptible, infected and recovered individuals, respectively, in population $i = 1, 2, \ldots, P$ at time $t \geq 0$. As the population size N is constant then $S_i(t) + I_i(t) + R_i(t) = N, \forall t \geq 0, i = 1, 2, \ldots, P$. The transition rates for the resulting $2P$-dimensional Markov chain from state $(s_1, i_1, s_2, i_2, \ldots, s_P, i_P)$ at time t are summarised in Table 4.1.

The metapopulation structure can be described by a weighted network $G = (V, E)$ with vertex set $V = \{1, 2, \ldots, P\}$ and edge set E, where edge $e = ij$ has weight σ_{ij}: the coupling matrix $\mathbf{\Sigma}$ therefore represents the weighted adjacency matrix for the graph G. For mathematical tractability we restrict our analysis to networks for which we can derive analytic results, namely graphs that are highly symmetric. In the following analysis we consider the complete network, the k-regular tree network and the star network. In addition, we assume that $\sigma_{ij} = \sigma, \forall ij \in E$. We note that for k-regular tree network and the star network, the weighted adjacency matrix $\mathbf{\Sigma}$ is sparse, that is, most of the elements are zero.

Throughout this chapter we will use the following notation for the first-order central

moments:

$$\begin{aligned}
\bar{S}_j &= \mathbb{E}[S_j] \\
\bar{I}_j &= \mathbb{E}[I_j] \\
C_{S_jS_j} &= \mathrm{Cov}(S_j, S_j) = \mathrm{Var}(S_j) \\
C_{I_jI_j} &= \mathrm{Cov}(I_j, I_j) = \mathrm{Var}(I_j) \\
C_{S_jI_j} &= \mathrm{Cov}(S_j, I_j),
\end{aligned}$$

and for the second-order moments:

$$\begin{aligned}
\hat{C}_{S_jS_k} &= \mathrm{Cov}(S_j, S_k) \\
\hat{C}_{I_jI_k} &= \mathrm{Cov}(I_j, I_k) \\
\hat{C}_{S_jI_k} &= \mathrm{Cov}(S_j, I_k),
\end{aligned}$$

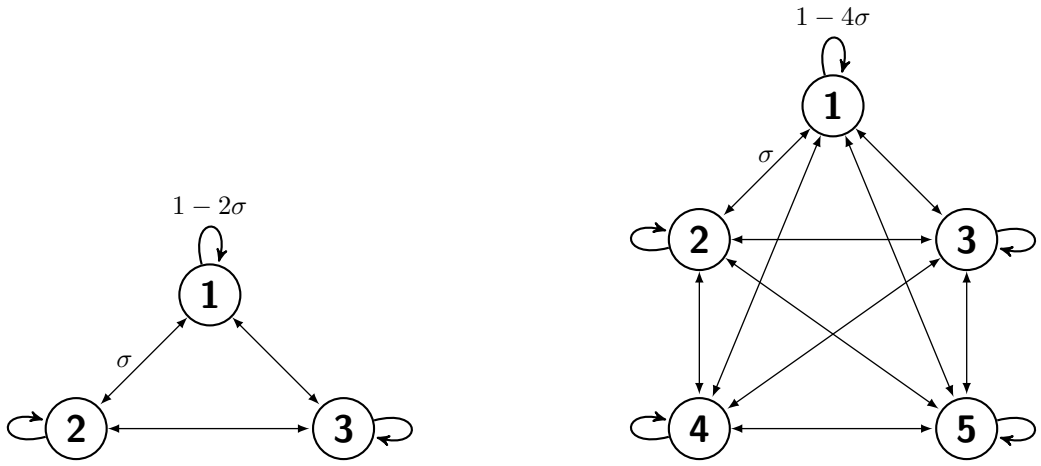
where X^* and C^*_{XY} denote the first- and second-order moments X and C_{XY} at endemic equilibrium, respectively.

For a metapopulation network on P populations, the set of ODEs approximating the stochastic process has at most $3P^2 + 2P$ equations: P for each of the two first order moments and P^2 for each of the three covariances. However, for the networks that we consider in this chapter, symmetries in the structure of the network mean that the effective number of equations at endemic equilibrium is reduced. In some cases we will simplify the notation: we outline simplifications to the notation at the start of the results section for each network.

4.3 The complete network

4.3.1 Network definition and notation

First we consider P identical populations on the complete network, where each population interacts with the other $k = P - 1$ populations: a visual representation of the complete network for $P = 3$ and $P = 5$ populations is given in Figure 4.1. The coupling matrix $\mathbf{\Sigma} = (\sigma_{ij})$ is defined as



(a) **(b)**

Figure 4.1. The complete network on (a) $P = 3$ and (b) $P = 5$ populations. The coupling between any pair of populations coupling is $\sigma \in [0, 1/(P-1)]$ and so the within-population coupling is $1 - (P-1)\sigma$.

$$\sigma_{ij} = \begin{cases} 1 - k\sigma, & \text{for } i = j \\ \sigma, & \text{for } i \neq j. \end{cases}$$

In the complete network metapopulation all subpopulations are epidemiologically and topologically identical: epidemiologically in the sense that all subpopulations are of equal size and have identical epidemiological parameters, and topologically in the sense that all nodes are isomorphic within the network and the coupling is the same between any pair of subpopulations. As a result, the expected behaviour is the same within all populations, and between any pair of populations. In our notation, we can therefore drop dependency on the population and simplify it to the following: $\bar{X} = \mathbb{E}[X_j]$, $C_{XY} = \mathrm{Cov}(X_j, Y_j)$ and $\hat{C}_{XY} = \mathrm{Cov}(X_i, Y_j), i \neq j$.

Using the second-order moment closure approximation, and with these simplifications, the stochastic process on the complete network can be approximated by a set of eight ODEs: five for the within-population moments, and three for the between-population moments. These can be found in Appendix B. We use these equations in both the analytical and the numerical results.

4.3.2 Analytic approximation for the correlation between any pair of subpopulations

For P populations on the complete network, we define the correlation between any pair of populations as

$$\rho = \frac{\hat{C}^*_{II}}{C^*_{II}},$$

and show that this is equal to

$$\rho = \frac{\sigma}{\xi + \sigma} - \Delta, \tag{4.1}$$

where

$$\xi = \frac{N(\gamma + \mu) - \beta \bar{S}^*}{\beta \bar{S}^*} \tag{4.2}$$

and

$$\Delta = \frac{(\beta \bar{I}^* + N\epsilon)\frac{\hat{C}^*_{SI}}{C^*_{II}}}{\beta(1-\sigma)\bar{S}^* - N(\gamma + \mu)}. \tag{4.3}$$

We derive this result by taking the moment equation for $\hat{C}_{II}$ at equilibrium and dividing through by $2C^*_{II}/N$, following the same approach as Chapter 2; full details of this derivation can be found in Appendix B. Moreover, if $\Delta \ll 1$ then we can further simplify the approximation for the correlation to the following expression:

$$\rho \approx \frac{\sigma}{\xi + \sigma}. \tag{4.4}$$

We can also use an alternative approximate expression for ξ that is independent of $\bar{S}^*$, which eliminates the need to find the equilibrium of the 8-dimensional ODE model.

In Chapter 2, we showed that by ignoring the effects of imports and correlations and taking the large population limit, then

$$\xi \approx \xi' = \frac{\epsilon(\gamma+\mu)}{\mu(\beta-\gamma-\mu)} = \frac{\epsilon}{\mu(R_0-1)}. \tag{4.5}$$

Given the simpler form of Equation (4.5) compared to the original expression for ξ given by Equation (4.2), in the remainder of the analysis we evaluate $\sigma/(\xi'+\sigma)$ as an approximation for the MVN correlation ρ.

This result is independent of the number of populations P. In short, this is due to the balance between two competing influences: the addition of an extra external coupling would normally weaken the correlation between two connected populations, but the fact that this additional population is itself correlated with the original populations nullifies this effect. At the end of the chapter, we make this argument explicit by adding a third population (with variable coupling) to an interacting pair of populations.

4.3.3 Numerical results

We look at the effect of the number of subpopulations P on the dynamics of the first- and second-order moments, and on the correlation. We use parameters representing a measles-like endemic disease in the UK: $N = 10^5$ where $R_0 = 17$, $\gamma^{-1} = 13$ days, $\mu = 5.5 \times 10^{-5}$ days^{-1} and $\epsilon = 5.5 \times 10^{-5}$ days^{-1}.

Dynamics of first- and second- order moments

We first explore the effect of the number of subpopulations P and coupling σ on the equilibrium values of the first-order central moments $\bar{S}^*$ and $\bar{I}^*$ and the second-order central moments C^*_{II} and $\hat{C}^*_{II}$ (Figure 4.2a). We consider $P = 3, 5, 10$ and $\sigma \in [0, 1/k], k = P-1$, and include $P = 2$ for comparison. These results are obtained by the numerical integration of the system of ODEs given in the Appendix B, and so only introduce an error due to the MVN moment closure approximation. For all values of P, all curves show a sigmoidal pattern, with $\bar{S}^*$ and C^*_{II} decreasing with the coupling, and $\bar{I}^*$ and $\hat{C}^*_{II}$ increasing with the coupling. As the number of populations P increases the magnitude of change in C^*_{II} increases, since reducing the within-population coupling (either by increasing the between-population coupling σ or increasing the number of populations P)

reduces the variance C_{II}. However, the magnitude of change in $\hat{C}^*_{II}$ decreases, because as P increases, then the effect of interaction between a subpopulation and its neighbour is damped by the other $P-2$ neighbours. In the previous section we noted that our approximation for the correlation is independent of the number of populations P: we also calculate the MVN correlation $\hat{C}^*_{II}/C^*_{II}$ (Figure 4.2b) and note that this also appears independent of P. The correlation follows a sigmoidal relationship, increasing from zero for very low coupling.

Comparison of the approximation and simulations

Next we compare the MVN correlation ρ (Equation (B.9)) and our approximation $\sigma/(\xi' + \sigma), \xi' = 0.0625$ (Equation (4.4)) to stochastic simulations for $P = 3, 5$ subpopulations (Figure 4.3). The close agreement between ρ and the simulation results suggests that our use of the MVN moment closure approximation is justified. There is also little difference between the MVN correlation and our approximation (that is, Δ is small), so $\sigma/(\xi' + \sigma)$ is a good approximation for the correlation ρ. Therefore, we can relate the phenomenological coupling parameter σ to the correlation between the number of infected individuals in any pair of populations for P populations arranged on the complete network by $\rho \approx \sigma/(\xi' + \sigma)$.

4.3.4 Independence of the number of subpopulations, P

The analytic approximation for the MVN correlation between two subpopulations on the complete network is independent of the number of populations P. In short, this is due to the balance between two competing influences: the addition of an extra external coupling would normally weaken the correlation between two connected populations, but the fact that this additional population is itself correlated with the original populations nullifies this effect. In the following section, we make this argument explicit by adding a third population (with variable coupling) to an interacting pair of populations.

We consider a metapopulation network on three populations with coupling matrix

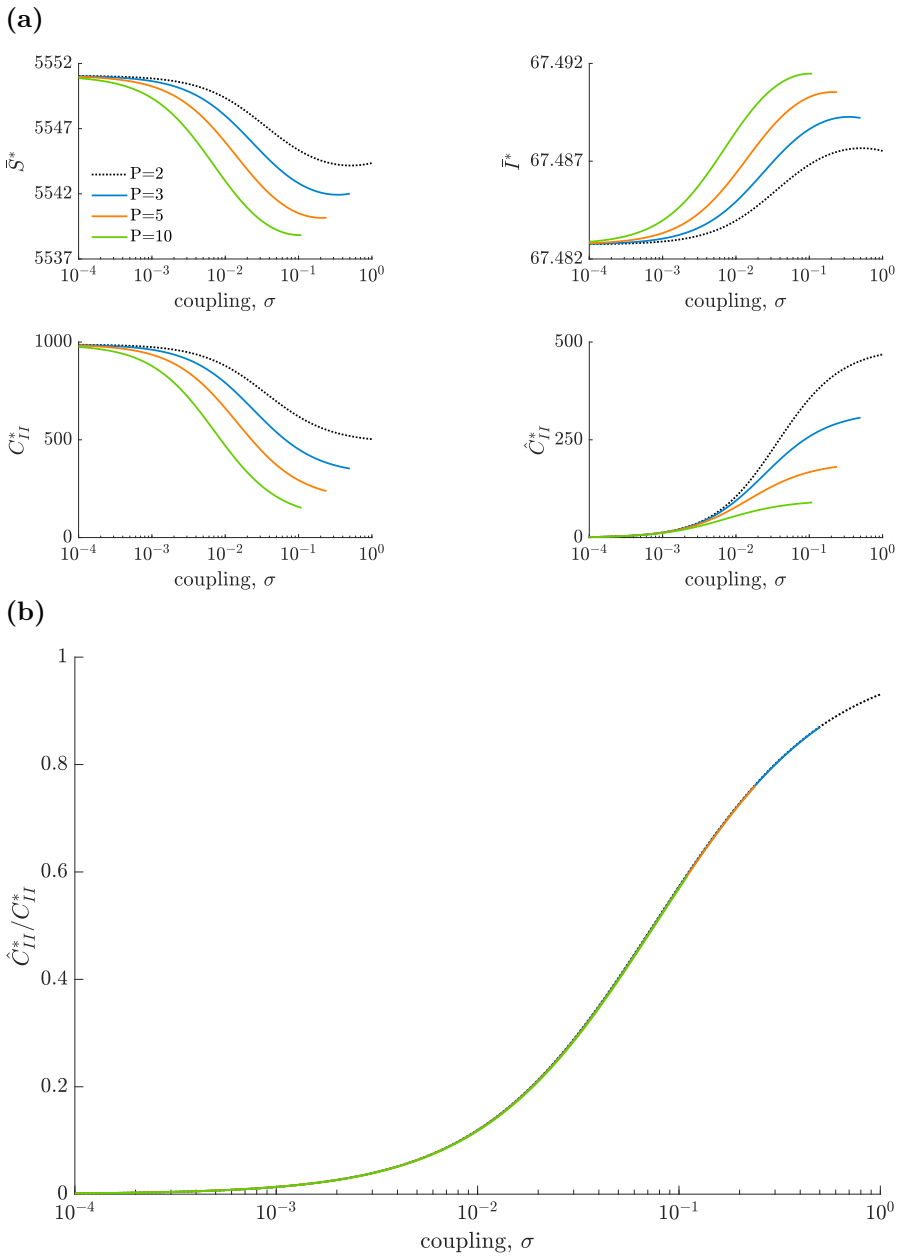


Figure 4.2. The effect of the coupling σ on (a) the key mean variables $\bar{S}^*, \bar{I}^*, C^*_{II}$ and $\hat{C}^*_{II}$ and (b) the correlation $\hat{C}^*_{II}/C^*_{II}$for P populations arranged on the complete network. Parameter values represent a measles-like endemic disease in the UK ($N = 10^5, \mu = 5.5 \times 10^{-5}, R_0 = 17, \gamma^{-1} = 13$ and $\epsilon = 5.5 \times 10^{-5}$). These values are calculated from the system of ODEs given in the Appendix B.

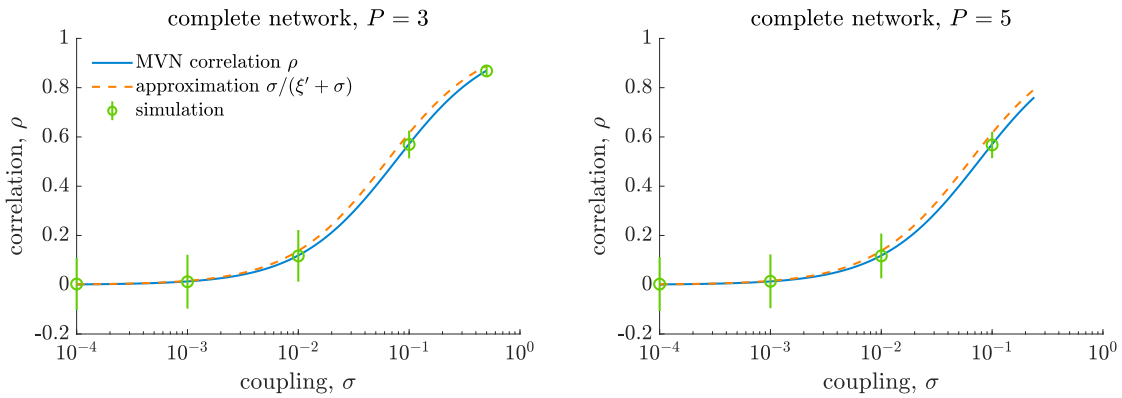


Figure 4.3. Comparing analytic and numerical correlation between any pair of populations from $P = 3, 5$ populations arranged on the complete network. We compare the analytic correlation ρ and our approximation $\sigma/(\xi' + \sigma), \xi' = 0.0625$, to stochastic simulations for a measles-like endemic disease in the UK ($N = 10^5, \mu = 5.5 \times 10^{-5}, R_0 = 17, \gamma^{-1} = 13$ and $\epsilon = 5.5 \times 10^{-5}$). Each population is coupled to the $k = P - 1$ other populations. The between-population coupling is fixed as $\sigma \in [0, 1/k]$ and within-population coupling is therefore $1 - k\sigma$. We generate 1000 realisations of the process for each value of σ and calculate the correlation as a time-weighted Pearson correlation coefficient for $50 \leq t \leq 200$; error bars represent ± 2 standard deviations.

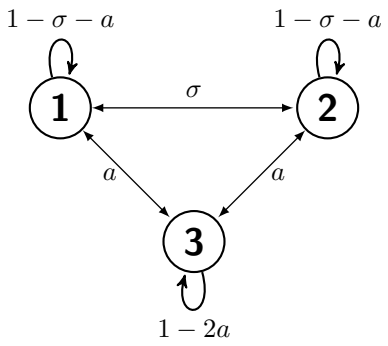


Figure 4.4. A metapopulation on three identical subpopulations. The coupling between populations 1 and 2 is $\sigma \in [0,1]$, and the coupling between populations 1 and 3, or 2 and 3, is $a \in [0,\sigma]$.

$\mathbf{\Sigma} = (\sigma_{ij})_{ij}$ defined as

$$\sigma_{ij} = \begin{cases} 1-\sigma-a, & \text{for } i=j=1,2 \\ 1-2a, & \text{for } i=j=3 \\ \sigma, & \text{for } (i,j) \in \{(1,2),(2,1)\} \\ a, & \text{for } (i,j) \in \{(1,3),(3,1),(2,3),(3,2)\}, \end{cases} \tag{4.6}$$

where $\sigma \in [0,1]$ and $a \in [0,\sigma]$. A visual representation of this metapopulation is given in Figure 4.4.

In this metapopulation all subpopulations are epidemiologically identical (that is, they are the same size and have identical epidemiological parameters). In addition, subpopulations 1 and 2 are isomorphic in the weighted network, but are only isomorphic with subpopulation 3 when $a = \sigma$. We retain our original notation, using $\bar{X}_i = \mathbb{E}[X_i]$, $C_{X_iY_i} = \mathrm{Cov}(X_i, Y_i)$ and $\hat{C}_{X_iY_j} = \mathrm{Cov}(X_iY_j)$.

Approximation for correlation between subpopulations 1 and 2

The correlation between subpopulations 1 and 2 is given by

$$\rho_{12} = \frac{\hat{C}_{I_1I_2}}{\sqrt{C_{I_1I_1}C_{I_2I_2}}} = \frac{\hat{C}_{I_1I_2}}{C_{I_2I_2}}. \tag{4.7}$$

To derive an approximation for this, we begin with the moment equation for $\hat{C}_{I_1I_2}$:

$$\frac{d\hat{C}_{I_1I_2}}{dt} = 2\Bigg[\left(\frac{\beta}{N}(1-\sigma-a)\bar{S}_1 - (\gamma+\mu)\right)\hat{C}_{I_1I_2} + \frac{\beta}{N}\sigma\bar{S}_1 C_{I_2I_2} + \frac{\beta}{N}a\bar{S}_1\hat{C}_{I_2I_3} + \left(\frac{\beta}{N}(1-\sigma-a)\bar{I}_1 + \frac{\beta}{N}\sigma\bar{I}_2 + \frac{\beta}{N}a\bar{I}_3 + \epsilon\right)\hat{C}_{S_1I_2}\Bigg]. \tag{4.8}$$

At equilibrium $d\hat{C}_{I_1I_2}/dt = 0$, and if we divide by $2C_{I_2I_2}/N$ then

$$\begin{aligned} 0 &= [\beta(1-\sigma-a)\bar{S}_1 - N(\gamma+\mu)]\rho_{12} + \beta\sigma\bar{S}_1 + \beta a\bar{S}_1\sqrt{\frac{V_3}{V_1}}\rho_{23} \\ &\quad + \left(\beta(1-\sigma-a)\bar{I}_1 + \beta\sigma\bar{I}_2 + \beta a\bar{I}_3 + \epsilon\right)\frac{\hat{C}_{S_1I_2}}{C_{I_2I_2}} \qquad (4.9) \\ &= [\beta(1-\sigma)\bar{S}_1 - N(\gamma+\mu)]\rho_{12} + \beta\sigma\bar{S}_1 + \beta a\bar{S}_1\left(\sqrt{\frac{V_3}{V_1}}\rho_{23} - \rho_{12}\right) \\ &\quad + \left(\beta(1-\sigma-a)\bar{I}_1 + \beta\sigma\bar{I}_2 + \beta a\bar{I}_3 + \epsilon\right)\frac{\hat{C}_{S_1I_2}}{C_{I_2I_2}}, \qquad (4.10) \end{aligned}$$

and hence we have the following approximation for the correlation between subpopulations 1 and 2:

$$\begin{aligned} \rho_{12} &= \frac{\beta\sigma\bar{S}_1}{N(\gamma+\mu) - \beta(1-\sigma)\bar{S}_1} + \frac{\beta a\bar{S}_1}{N(\gamma+\mu) - \beta(1-\sigma)\bar{S}_1}\left(\sqrt{\frac{V_3}{V_1}}\rho_{23} - \rho_{12}\right) - \Delta \qquad (4.11) \\ &= \frac{\sigma}{\xi+\sigma} + \frac{a}{\xi+\sigma}\left(\sqrt{\frac{V_3}{V_1}}\rho_{23} - \rho_{12}\right) - \Delta, \qquad (4.12) \end{aligned}$$

where Δ is the correction term given by:

$$\Delta = \frac{(\beta(1-\sigma-a)\bar{I}_1 + \beta\sigma\bar{I}_2 + \beta a\bar{I}_3 + \epsilon)}{\beta(1-\sigma)\bar{S}_1 - N(\gamma+\mu)}\frac{\hat{C}_{S_1I_2}}{C_{I_2I_2}}. \tag{4.13}$$

If $\Delta \ll 1$ and $\xi \approx \xi'$, we have the following simplified expression for the correlation:

$$\rho_{12} \approx \frac{\sigma}{\xi'+\sigma} + \frac{a}{\xi'+\sigma}\left(\sqrt{\frac{V_3}{V_1}}\rho_{23} - \rho_{12}\right). \tag{4.14}$$

We consider how the second term of Equation (4.14) changes for $a \in [0, \sigma]$. When $a = 0$ then the network becomes the complete network on two subpopulations, and in Equation (4.14), the second term vanishes and $\rho_{12} = \sigma/(\xi' + \sigma)$, as expected. When

$a = \sigma$ the network becomes the complete network on three subpopulations, and in Equation (4.14), $V_1 = V_3$ and $\rho_{12} = \rho_{23}$, so $(\sqrt{V_3/V_1}\rho_{23} - \rho_{12}) = 0$ and so we also have $\rho_{12} = \sigma/(\xi' + \sigma)$. If $a \in (0, \sigma)$ then $0 < \rho_{23} < \rho_{12}$. Therefore, if $\sqrt{V3/V1} < \rho_{12}/\rho_{23}$ then $(\sqrt{V_3/V_1}\rho_{23} - \rho_{12}) < 0$ and $\rho_{12} < \sigma/(\xi' + \sigma)$.

4.4 The tree network

4.4.1 Network definition and notation

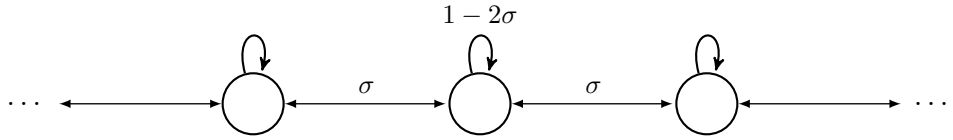
Next, we consider infinitely many populations on a k-regular tree network, where each subpopulation has k neighbours: a visualisation of the k-regular tree network for $k = 2$ and $k = 4$ neighbours is given in Figure 4.5. The coupling matrix $\mathbf{\Sigma} = (\sigma_{ij})$ is defined as

$$\sigma_{ij} = \begin{cases} 1 - k\sigma, & \text{for } i = j \\ \sigma, & \text{for } i, j \text{ neighbours, } i \neq j \\ 0, & \text{otherwise.} \end{cases} \tag{4.15}$$

As with the complete network, all subpopulations in the k-regular tree network are epidemiologically and topologically identical, so the expected behaviour is the same within all subpopulations. In addition, in a tree network, there is a unique path between any pair of subpopulations, and so we can define the distance $d_{ij} \in \mathbb{N}$ between subpopulations i and j to be the length of the path between the subpopulations. For the notation for within-population moments we can again drop dependency on the subpopulation: $\bar{X} = \mathbb{E}[X_j]$ and $C_{XY} = \text{Cov}(X_j, Y_j)$. For the between-population moments, we only need to denote the distance d between the subpopulations: $\hat{C}^{(d)}_{XY} = \text{Cov}(X_i, Y_j), i \neq j$, where $d_{ij} = d$.

Using the second-order moment closure approximation, we can write ODEs for the first and second-order moments of the stochastic process on the k-regular tree network. However, this system comprises infinitely many equations: five equations for the within-population moments, and infinitely many equations for the between-population moments (3 for each $d \geq 1$). In addition, we cannot perform stochastic simulations of the infection process on infinitely many subpopulations. To overcome these problems, we consider a

(a)



(b)

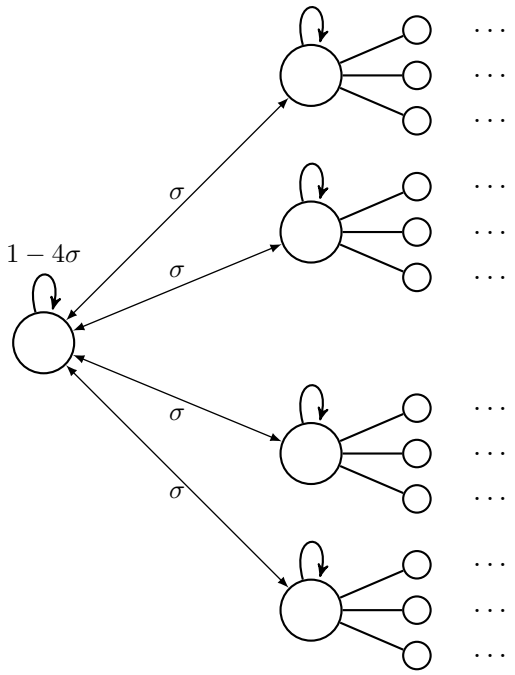


Figure 4.5. The k-regular tree network for (a) $k = 2$ and (b) $k = 4$ neighbours. The coupling between any pair of neighbouring populations is $\sigma \in [0, 1/k]$ and so the within-population coupling is $1 - k\sigma$.

finite subgraph approximation of the k-regular tree network.

Finite subgraph approximation of the k-regular tree network

We consider a finite subgraph of the k-regular tree network: we define the D-truncated k-regular tree network to be the network of subpopulations distance less than or equal to D from some arbitrarily chosen origin node; since all subpopulations are identical and the k-regular tree network is infinite, the choice of origin node is irrelevant. The total number of subpopulations in the D-truncated k-regular tree network is

$$T = 1 + k \sum_{i=0}^{D-1} (k-1)^i. \tag{4.16}$$

If D is sufficiently large, then we can make some further simplifying assumptions about the first- and second-order moments of the stochastic process on the D-truncated k-regular tree network.

First, we assume that covariances are negligibly small for subpopulations that are far apart: $\hat{C}_{XY}^{(d)} = 0, \forall d > D$. As a result, we can write down a smaller and simpler set of ODEs that approximate the stochastic process on the D-truncated k-regular tree network: without this assumption, we would also have to write down ODEs for covariances $\hat{C}_{XY}^{(d)}$ where $D < d \leq 2D$. Second, we maintain the simplification from the full k-regular tree network that covariances between any pair of subpopulations the same distance apart are the same. Although this is not true (the covariances between adjacent populations at the centre of the truncated tree network will be different to covariances between adjacent subpopulations at the edge of the network, for example), in practice it will have little effect on the final results. This assumption also considerably reduces the total number of ODEs we need to write down to approximate the stochastic process. Third, we assume that the expected behaviour of the first- and second-order central moments in the origin node, and between the origin node and subpopulations at distance $d \ll D$ will be the same as in the full k-regular tree network. In this way, results derived for the D-truncated k-regular tree network will be the same as the results in the full k-regular tree network.

Given these assumptions, and making a second-order MVN moment closure approximation, the stochastic process on the D-truncated k-regular tree network can be approximated by a set of $5 + 3D$ equations: five equations for the within-population moments and $3D$ equations for the between-population moments. These can be found in Appendix

B.

4.4.2 Analytic approximation for the correlation between subpopulations distance d apart

We can derive analytic results for the full k-regular tree network rather than the finite subgraph approximation. We define the correlation between the number of infected individuals in a pair of subpopulations distance $d \geq 1$ apart as

$$\rho_d = \frac{\hat{C}_{II}^{(d)*}}{C_{II}^*},$$

where $\rho_0 = 1$ and $\lim_{d\to\infty} \rho_d = 0$. We can show that ρ_d is the solution to

$$\rho_d = \frac{\sigma}{\xi + k\sigma} \left(\rho_{d-1} + (k-1)\rho_{d+1}\right) - \Delta^{(d)}, \tag{4.17}$$

where

$$\xi = \frac{N(\gamma + \mu) - \beta \bar{S}^*}{\beta \bar{S}^*} \tag{4.18}$$

and

$$\Delta_k^{(d)} = \frac{(\beta \bar{I}^* + N\epsilon)}{\beta(1 - k\sigma)\bar{S}^* - N(\gamma + \mu)} \frac{\hat{C}_{SI}^{(d)*}}{C_{II}^*}. \tag{4.19}$$

We derive this result from the moment equation for $\hat{C}_{II}^{(1)}$ at equilibrium and dividing through by $2C_{II}^*/N$; full details of this derivation can be found in Appendix B. Moreover, if $\Delta^{(d)} \ll 1, \forall d$ then ρ_d is the solution to the recurrence relation

$$(k-1)\rho_{d+1} = \frac{\xi + k\sigma}{\sigma}\rho_d - \rho_{d-1}, \tag{4.20}$$

where $\rho_0 = 1$ and $\lim_{d\to\infty} \rho_d = 0$. Since $|\rho_d| \leq 1$ then the solution is given by

$$\rho_d = \left(\frac{k\sigma + \xi - \sqrt{\xi^2 + 2k\xi\sigma + (k-2)^2\sigma^2}}{2(k-1)\sigma} \right)^d$$
$$= \left(\frac{k\sigma + \xi - \sqrt{\sigma^2 k^2 + (2\xi\sigma - 4\sigma^2)k + 4\sigma^2 + \xi^2}}{2(k-1)\sigma} \right)^d . \quad (4.21)$$

We note two things: firstly, since $\rho_1 \leq 1$ then it is trivial that $\rho_d \to 0$ as $d \to \infty$. Secondly, $\rho_d \to 0$ as $k \to \infty$.

4.4.3 Numerical results

As noted earlier, numerical analysis must be conducted on the D-truncated k-regular tree network: stochastic simulations and the MVN correlation are evaluated on this finite subgraph approximation as it is not possible to use the full k-regular tree network. If D is sufficiently large, then these correlations will be approximately the same as in the full k-regular tree network.

We look at the effect of the number of neighbours k on the correlation. Again, we use parameters representing a measles-like endemic disease in the UK: $N = 10^5$ where $R_0 = 17$, $\gamma^{-1} = 13$ days, $\mu = 5.5 \times 10^{-5}$ days^{-1} and $\epsilon = 5.5 \times 10^{-5}$ days^{-1}.

Effect of number of neighbours k and distance d on the MVN correlation

We first numerically evaluate the effect of the number of neighbouring subpopulations k and the distance d on the correlation ρ_d (Figure 4.6). As with the complete network, the correlation follows a sigmoidal shape, increasing from zero correlation from very low coupling. For fixed coupling σ, as the number of neighbours k increases then the correlation ρ_d decreases; similarly, for a fixed number of neighbours k, as the distance d increases then the correlation ρ_d also decreases. This all agrees with expected behaviour from Equation (4.21).

Effect of D on the correlation

As D increases, the total number of subpopulations in the D-truncated k-regular tree network increases rapidly, which significantly impacts the speed at which simulations

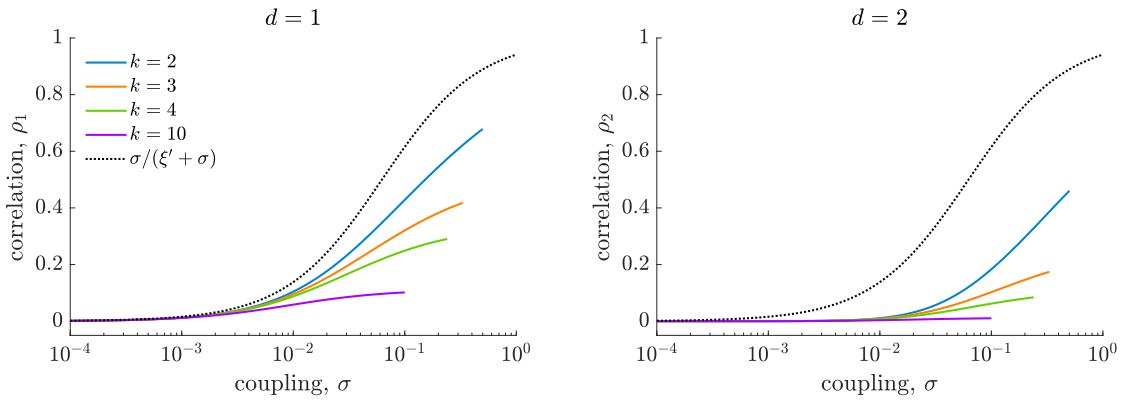


Figure 4.6. The effect of the number of neighbouring subpopulations k in the k-regular tree network on the correlation between the number of infected individuals in adjacent populations, ρ_1 (left), and populations with a common neighbour, ρ_2 (right). Parameter values represent a measles-like endemic disease in the UK ($N = 10^5$, $\mu = 5.5 \times 10^{-5}$, $R_0 = 17$, $\epsilon = 5.5 \times 10^{-5}$, $\gamma = 1/13$). The MVN correlation is calculated on the D-truncated k-regular tree network for $D = 50$ from the system of ODEs given in Appendix B.

can be performed. In this section we consider the effect of D on the correlation.

We calculate the correlation ρ_1 between adjacent populations in the D-truncated k-regular tree network for $k = 2, 4$ and a range of values for D (Figure 4.7). For low coupling ($\sigma = 0.001, 0.01$), increasing D has little effect on the correlation: this suggests that when coupling is low it is sufficient to take $D = 3$. However, for larger coupling ($\sigma = 0.1$) then as D increases then the correlation initially decreases, and then plateaus: for both $k = 2$ and $k = 4$ the correlation plateaus around $D = 5$.

Comparison of the approximation and simulations

Next, we compare our approximations to the results of stochastic simulations for $k = 2, 4$ (Figure 4.8), where stochastic simulations are performed on the D-truncated k-regular tree network and $D = 5, 3$ for $k = 2, 4$, respectively. For all combinations of k and d there is close agreement between the MVN correlation and stochastic simulations, which justifies our use of the MVN moment closure approximation, and we showed in the previous section that increasing D further does not significantly change the correlations in the system (Figure 4.7). There is also little difference between the MVN correlation and our approximation (that is, $\Delta^{(1)}$ is small) and so approximating the MVN correlation

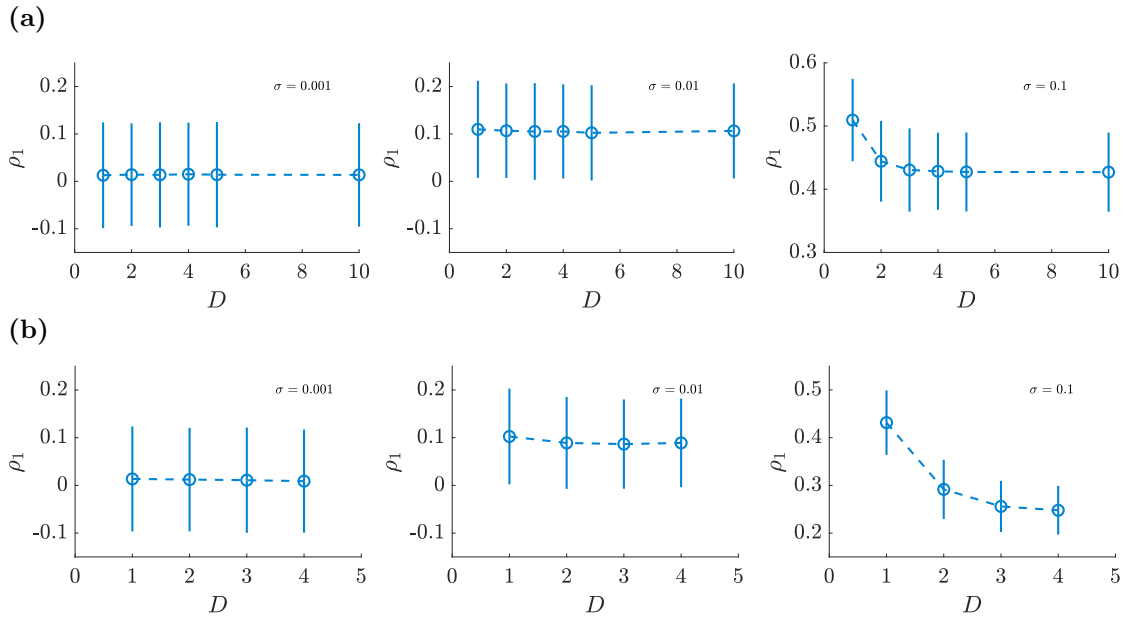


Figure 4.7. The correlation ρ_1 between adjacent subpopulations in the D-truncated k-regular tree network for (a) $k = 2$, and (b) $k = 4$. Parameter values represent a measles-like endemic disease in the UK ($N = 10^5, \mu = 5.5 \times 10^{-5}, R_0 = 17, \gamma^{-1} = 13$ and $\epsilon = 5.5 \times 10^{-5}$). We generate 1000 realisations of the process for each value of σ and calculate the correlation as a time-weighted Pearson correlation coefficient for $50 \leq t \leq 200$; error bars represent ± 2 standard deviations.

by Equation (4.21) is reasonable.

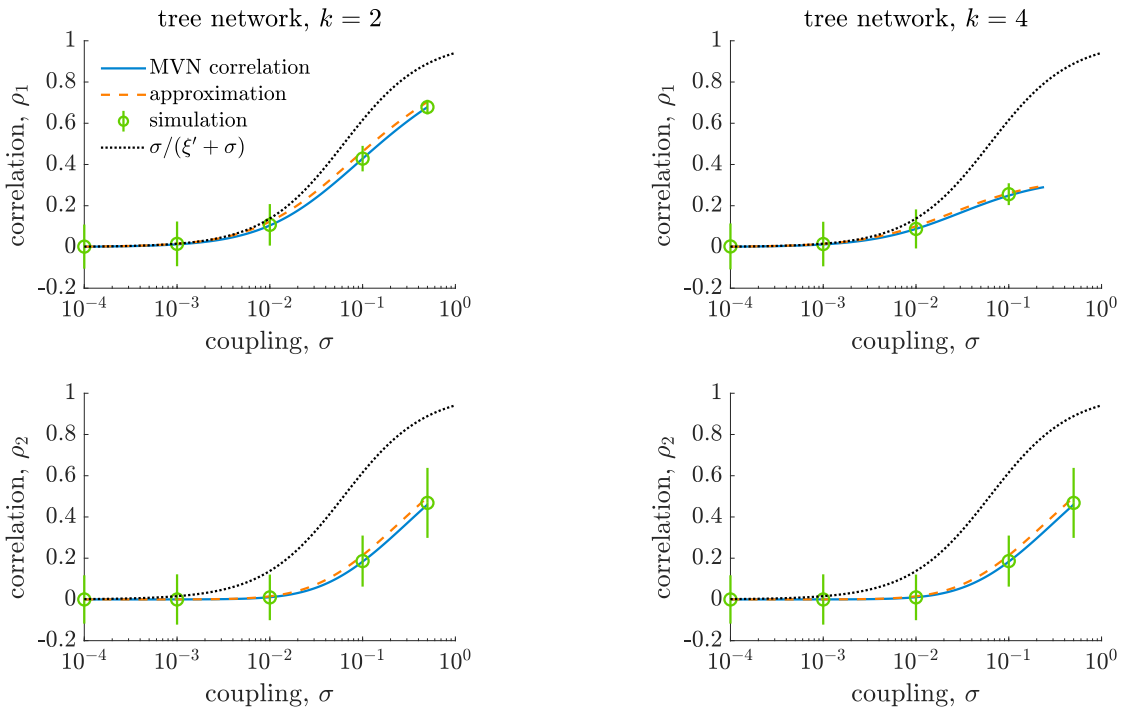


Figure 4.8. Comparing the MVN correlation ρ_d and our approximation to stochastic simulations for a measles-like endemic disease in the UK in T populations arranged on the D-truncated k-regular tree network ($N = 10^5, \mu = 5.5 \times 10^{-5}, \beta = 17/13, \epsilon = 5.5 \times 10^{-5}, \gamma = 1/13$). The coupling between interacting populations is $\sigma \in [0, 1/k]$. The stochastic process is simulated on the D-truncated k-regular tree network, with $D = 5$ and $D = 3$ for $k = 2, 4$, respectively. The process is simulated over a 200 year period using the Gillespie algorithm, with a burn-in period of 50 years, and generate 100 realisations of the process for each value of σ. The correlation is calculated as a time-weighted Pearson correlation coefficient for $50 \leq t \leq 200$; error bars represent ± 2 standard deviations.

4.5 The star network

Network definition and notation

Finally, we consider the star network on P subpopulations, where there is a central 'hub' subpopulation (labelled as subpopulation 1) and $k = P - 1$ 'leaf' populations; there is no

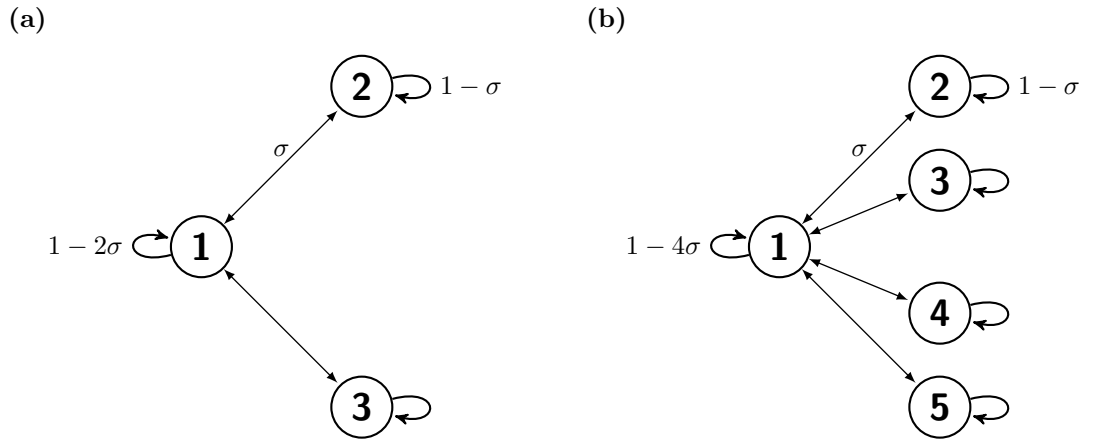


Figure 4.9. The star network on (a) $P = 3$ and (b) $P = 5$ populations. The coupling between any pair of neighbouring populations is $\sigma \in [0, 1/(P-1)]$ and so the within-population coupling is $1 - (P-1)\sigma$ for the hub population and $1 - \sigma$ for any leaf population.

direct interaction between the leaf populations. A visualisation of the star network for $P = 3$ and $P = 5$ subpopulations is given in Figure 4.9. The coupling matrix $\mathbf{\Sigma} = (\sigma_{ij})$ is defined as

$$\sigma_{ij} = \begin{cases} 1 - k\sigma, & \text{for } i = j = 1 \\ 1 - \sigma, & \text{for } i = j \neq 1 \\ \sigma, & \text{for } i = 1, j \neq 1 \text{ and } i \neq 1, j = 1 \\ 0, & \text{otherwise.} \end{cases}$$

Unlike the complete network and the k-regular tree network, the expected behaviour of the stochastic process is not the same within and between all subpopulations. This is because the hub subpopulation has k neighbours, whereas each leaf subpopulation has only one neighbour. However, we can still make some simplifications to the notation: the expected behaviour of the infection process is the same within any leaf subpopulation, or between any pair of leaf subpopulations, or between a leaf subpopulation and the hub subpopulation. We can therefore simplify our notation to distinguish between hub and leaf subpopulations. For the within-population moments, the sub/superscript indicates whether the subpopulation is a hub (H) or a leaf (L) subpopulation:

$$\begin{aligned}
\bar{X}_H &= \mathbb{E}[X_1] \\
\bar{X}_L &= \mathbb{E}[X_i], \quad i = 2, \ldots, P \\
C^H_{XY} &= \mathrm{cov}(X_1, Y_1) \\
C^L_{XY} &= \mathrm{cov}(X_i, Y_i), \quad i = 2, \ldots, P.
\end{aligned}$$

For the between-population moments, the sub/superscript indicates whether one of the subpopulations is a hub (H) or if they are both leaf subpopulations (L); for $\hat{C}_{S_iI_j}$ we distinguish between $\hat{C}_{S_1I_i}$ and $\hat{C}_{S_iI_1}$:

$$\begin{aligned}
\hat{C}^H_{XX} &= \mathrm{cov}(X_1, X_i), \quad i = 2, \ldots, P \\
\hat{C}^L_{XX} &= \mathrm{cov}(X_i, X_j), \quad i, j = 2, \ldots, P, i \neq j \\
\hat{C}_{X_HY_L} &= \mathrm{cov}(X_1, Y_i), \quad i = 2, \ldots, P.
\end{aligned}$$

Using the second-order moment closure approximation, the stochastic process on the star network for P subpopulations can be approximated by a set of seventeen ODEs: ten equations for the within-population moments, and seven equations for the between-population moments. These can be found in Appendix B. We use these equations in both the analytical and the numerical results.

4.5.1 Analytic results

For P identical subpopulations on the star network, we define the correlation between the number of infected individuals in the hub population and the number of infected individuals in a leaf population as

$$\rho_H = \frac{\hat{C}^{H*}_{II}}{\sqrt{C^{H*}_{II} C^{L*}_{II}}},$$

and the correlation between the number of infected individuals in two leaf subpopulations as

$$\rho_L = \frac{\hat{C}_{II}^{L*}}{C_{II}^{L*}}.$$

We can show that ρ_H and ρ_L are solutions to the following pair of simultaneous equations:

$$\rho_H = \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}} \frac{\sigma}{\frac{S_H^*}{S_L^*}(\xi_H + k\sigma) + \xi_L + \sigma} + \sqrt{\frac{C_{II}^{L*}}{C_{II}^{H*}}} \frac{\sigma}{\xi_H + k\sigma + \frac{S_L^*}{S_H^*}(\xi_L + \sigma)}(1 - (k-1)\rho_L) + \Delta_H \tag{4.22}$$

$$\rho_L = \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}} \frac{\sigma}{\xi_L + \sigma} \rho_H + \Delta_L, \tag{4.23}$$

where

$$\xi_H = \frac{N(\gamma + \mu) - \beta \bar{S}_H^*}{\beta \bar{S}_H^*}, \tag{4.24}$$

$$\xi_L = \frac{N(\gamma + \mu) - \beta \bar{S}_L^*}{\beta \bar{S}_L^*} \tag{4.25}$$

and

$$\begin{aligned} \Delta_H = &\frac{\beta(1 - k\sigma)\bar{I}_H^* + k\beta\sigma\bar{I}_L^* + N\epsilon}{2N(\gamma + \mu) - \beta(1 - k\sigma)\bar{S}_H^* - \beta(1 - \sigma)\bar{S}_L^*} \frac{\hat{C}_{S_H I_L}}{\sqrt{C_{II}^{H*} C_{II}^{L*}}} \\ &+ \frac{\beta(1 - \sigma)\bar{I}_L^* + \beta\sigma\bar{I}_H^* + N\epsilon}{2N(\gamma + \mu) - \beta(1 - k\sigma)\bar{S}_H^* - \beta(1 - \sigma)\bar{S}_L^*} \frac{\hat{C}_{S_L I_H}}{\sqrt{C_{II}^{H*} C_{II}^{L*}}} \end{aligned} \tag{4.26}$$

$$\Delta_L = \frac{\beta(1 - \sigma)\bar{I}_L^* + \beta\sigma\bar{I}_H^* + N\epsilon}{N(\gamma + \mu) - \beta(1 - \sigma)\bar{S}_L^*} \frac{\hat{C}_{SI}^L}{C_{II}^L}. \tag{4.27}$$

We derive this result by taking the moment equation for $\hat{C}_{II}^H$ and $\hat{C}_{II}^L$ at equilibrium; full details of this derivation can be found in Appendix B. Moreover, if $\Delta_H, \Delta_L \ll 1$ then we can further simplify this result to the following pair of simultaneous equations:

$$\rho_H \approx \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}} \frac{\sigma}{\frac{S_H^*}{S_L^*}(\xi_H + k\sigma) + \xi_L + \sigma} + \sqrt{\frac{C_{II}^{L*}}{C_{II}^{H*}}} \frac{\sigma}{\xi_H + k\sigma + \frac{S_L^*}{S_H^*}(\xi_L + \sigma)} (1 - (k-1)\rho_L) \tag{4.28}$$

$$\rho_L \approx \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}} \frac{\sigma}{\xi_L + \sigma} \rho_H. \tag{4.29}$$

4.5.2 Numerical results

We look at the effect of the number of subpopulations $k+1$ on the correlation. Again, we use parameters representing a measles-like endemic disease in the UK: $N = 10^5$ where $R_0 = 17$, $\gamma^{-1} = 13$ days, $\mu = 5.5 \times 10^{-5}$ days^{-1} and $\epsilon = 5.5 \times 10^{-5}$ days^{-1}.

Effect of the number of subpopulations $k+1$ on the MVN correlation

We first numerically evaluate the effect of the number of leaf subpopulations k on the correlations ρ_H and ρ_L (Figure 4.10). Firstly, we note that, as with the complete and tree network, both ρ_H and ρ_L exhibit a sigmoidal shape, increasing from zero correlation from very low coupling. Secondly, the correlation between two leaf nodes is lower than between the hub and a leaf node; this is to be expected, as the leaf nodes are not directly connected to each other. Finally for a given coupling σ as the number of neighbours k increases then the correlation decreases; this holds for both ρ_H and ρ_L.

Comparison of the approximation and simulations

We compare the MVN correlation and our approximation to the results of stochastic simulations (Figure 4.11). Firstly, we observe a close agreement between the MVN correlation and the stochastic simulations, which suggests that our use of the MVN moment closure approximation is justified. Secondly, there is little difference between the MVN correlation and our approximation (that is, Δ_H and Δ_L are small), and so our approximation is reasonable.

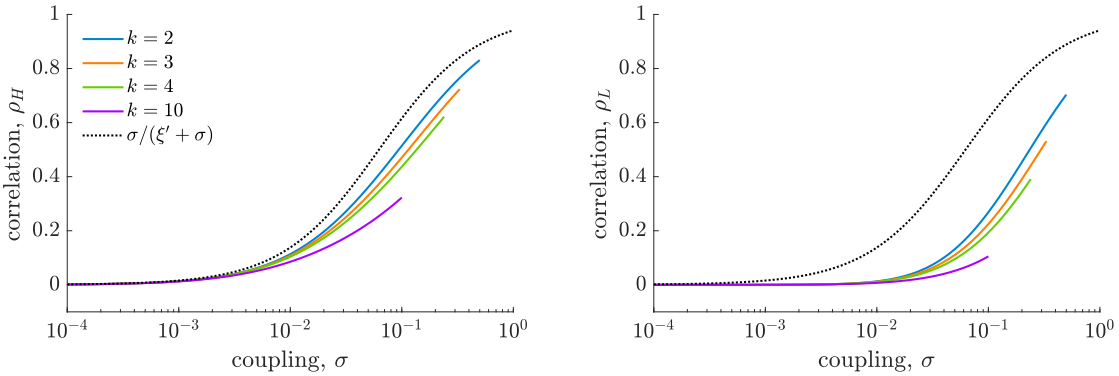


Figure 4.10. The effect of the number of leaf subpopulations k in the star network on the correlation between the number of infected individuals in the hub and a leaf population, ρ_H (left), and two leaf populations, ρ_L (right). Parameter values represent a measles-like endemic disease in the UK ($N = 10^5$, $\mu = 5.5\times10^{-5}$, $R_0 = 17$, $\epsilon = 5.5\times10^{-5}$, $\gamma = 1/13$). These values are calculated from the system of ODEs given in the Appendix B.

4.6 Comparison of networks

We now compare our approximations to the correlation between the number of infected individuals in adjacent subpopulations for all three networks (Figure 4.12). All networks are chosen to have the same k external connections: the complete network with $P = k+1$ populations, the k-regular tree network, and the star network with $P = k+1$ populations. We observe that the correlation is highest in the complete network and lowest in the tree network. Moreover, the difference between the approximations increases as k increases.

We attribute this behaviour to the total number of neighbour subpopulations that the two focal subpopulations have, how many of those neighbours are common neighbours, and whether these common neighbours interact. As the total number of neighbours of each member of the focal pair increases then the correlation decreases; for a given total number of neighbours the correlation is higher when more of these neighbours are common between the two focal subpopulations, and is higher yet when these common neighbours also interact with each other.

For a given k, two focal subpopulations in the complete network and the star network both have a total of $k - 1$ subpopulations. In the star network, none of these subpopulations are common neighbours of the two focal subpopulations; however, in the complete network, all these subpopulations are common neighbours and all the common

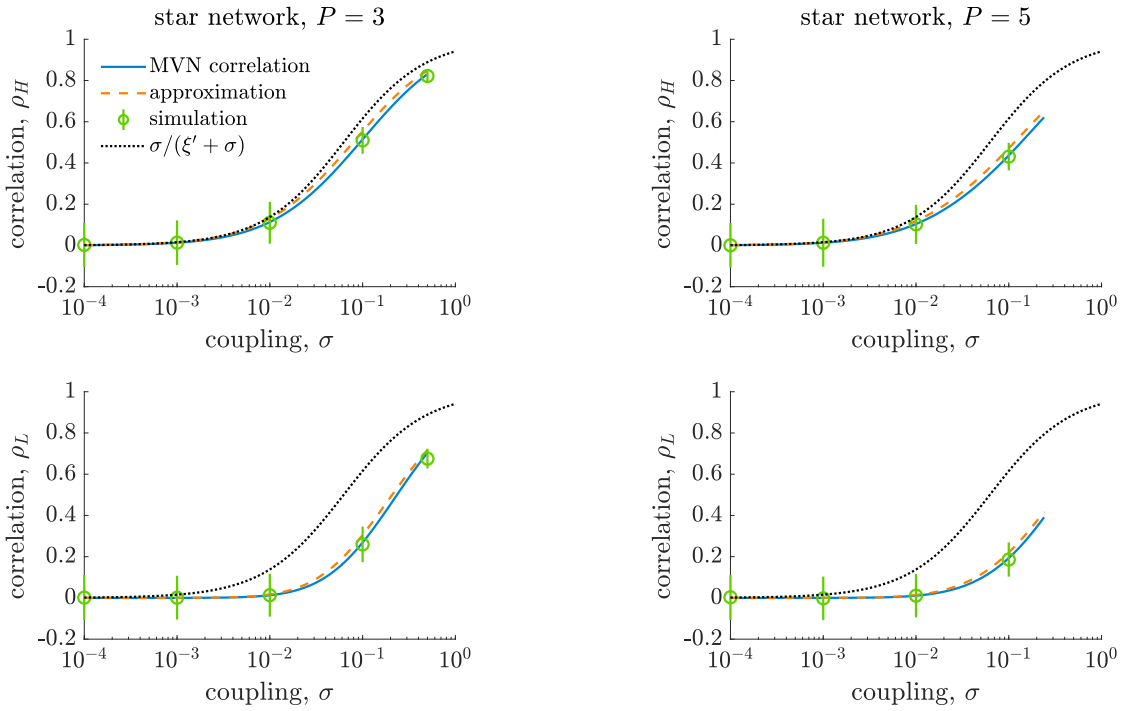


Figure 4.11. Comparing the analytic correlation, ρ_H and ρ_L, and our approximation to stochastic simulations for a measles-like endemic disease in the UK in $P+1$ populations arranged on the star network ($N = 10^5, \mu = 5.5 \times 10^{-5}, \beta = 17/13, \epsilon = 5.5 \times 10^{-5}, \gamma = 1/13$). The between-population coupling is fixed as $\sigma \in [0, 1]$ and within-population coupling is therefore $1 - \sigma$ in the hub population and $1 - \sigma$ in any leaf population. The stochastic process is simulated over a 200 year period using the Gillespie algorithm, with a burn-in period of 50 years, and generate 1000 realisations of the process for each value of σ. The correlation is calculated as a time-weighted Pearson correlation coefficient for $50 \leq t \leq 200$; error bars represent ± 2 standard deviations.

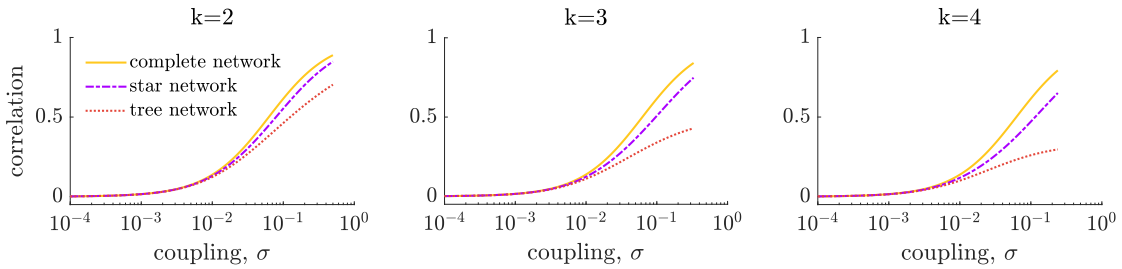


Figure 4.12. Comparison of our approximation to the correlation between a pair of adjacent populations in the complete network with $P = k+1$ populations, the k-regular tree network and the star network with $P = k+1$ populations.

neighbours interact with each other, hence the correlation in the star network is lower than in the complete network. For the same k, two focal subpopulations in the k-regular tree network have twice the total number of neighbours compared to the star network and none of these neighbours are common neighbours for either network. As a result, the correlation is lower in the tree network than in the star network.

4.7 Discussion

A limitation of metapopulation models in epidemiological modelling is now to infer the coupling between subpopulations: existing models do not accurately describe human mobility in developing countries, such as Sub-Saharan Africa, and sufficiently detailed data on human mobility are often lacking. This work provides insight into the effect of coupling and metapopulation network structure on endemic disease dynamics, and continues to develop a method for inferring the coupling between subpopulations in metapopulation models using disease prevalence data. We derive an approximation for the correlation ρ between the number of infected individuals in a given pair of subpopulations in certain network structures as a function of the coupling parameter σ. This provides a one-to-one mapping between the observable correlation ρ and the unknown coupling σ.

Our results extend the analysis of Chapter 2 from a simple two-population system to multiple populations arranged on a complete network, a k-regular tree network and a star network. Although we consider highly symmetric metapopulation networks, increased network complexity significantly reduces the analytic tractability of the model, compared to the two-population system. An alternative analytic relationship between the coupling

and correlation has previously been derived for more general networks (Rozhnova et al., 2012); however, we believe that our results provide greater intuition and analytical traction.

In addition, these results improve our understanding of how metapopulation network structure affects endemic disease dynamics in the metapopulation as a whole, complementing existing research on epidemic diseases in metapopulation networks (Barthélemy et al., 2010; Lahodny and Allen, 2013; Wang and Wu, 2018; Yan et al., 2018). We find that network distance between subpopulations and network structure are key drivers of the correlation, although, surprisingly, in the complete network the correlation between any pair of subpopulations is independent of the total number of subpopulations. We hypothesise that the correlation between two given subpopulations is driven by the number of neighbour subpopulations they both have, how many of these neighbours are shared between both subpopulations, and interactions between the neighbours.

We propose that disease prevalence data could be used to infer the underlying coupling from observed correlations between subpopulations in a metapopulation model. Our results provide insight into the effect of metapopulation network structure on endemic disease dynamics, but further work is required before it may be implemented in a real world setting. It would be useful to extend the results presented here to more realistic models of infectious disease dynamics, such as to include additional compartments or a seasonal component. This analysis could then be used to understand how the proposed method is affected by other mechanisms that contribute to temporally resolved correlation. The simple network structures we consider here do not fully capture the observed characteristics of real-world spatial networks, such as heterogeneous population size, degree or edge weight (Guimerà et al., 2005; Colizza et al., 2006). A natural extension of our current results is to allow heterogeneity in the epidemic parameters or metapopulation network structure, although we showed in Chapter 2, Section 2.3 that heterogeneous population sizes significantly impact the tractability of the results. In this case, a simulation-based study may be useful to determine how the correlation between two focal subpopulations is affected by their neighbours, their neighbours' neighbours and possible interactions between neighbours. This will allow us to elucidate which are the most important drivers of network correlations and overall endemic disease dynamics. There are additional practical questions that should be considered before it may be applied in a real world setting, such as: how much disease incidence data must be observed before accurate estimates of the correlation, and hence coupling, can be made; and whether the full metapopulation network structure needs to be known, as in reality

this is typically not the case. These extensions will move the results outlined here further towards a method for inferring coupling from correlations between subpopulations, thus addressing a key challenge of metapopulation modelling.

4.8 Conclusions

A limitation of metapopulation models in epidemiological modelling is how to infer the coupling between subpopulations. In this chapter we relate the correlation between the number of infected individuals in two populations as a function of the coupling, considering systems of multiple identical interacting populations on highly-symmetric complex networks. Our results provide insight into the effect of metapopulation network structure on endemic disease dynamics and provides the next step in developing a method for inferring coupling between subpopulations using disease prevalence data.

Chapter 5

Correlations between stochastic endemic infection in a general metapopulation network

5.1 Introduction

In Chapters 2 and 4, we derived analytic expressions for the correlation between prevalence of infection in two subpopulations in simple and highly symmetric metapopulation networks, using a multivariate normal moment closure assumption. It is challenging to extend this method to general metapopulation networks: we need to derive a system of simultaneous equations that can be solved for the pairwise correlations, but this is time-consuming to do by hand and not straightforward even using symbolic programming packages.

In this chapter we present an alternative method that allows us to numerically estimate the correlation between the prevalence of infection in two subpopulations in a general metapopulation network. By approximating the continuous-time Markov model of endemic disease dynamics by a diffusion process and making some additional simplifying assumptions, we can estimate the correlation between any pair of subpopulations.

Using this method we are able to study how the metapopulation network structure affects the correlation between subpopulations. We consider multiple network configurations, ranging from small metapopulations with $P = 4$ subpopulations, to generalised star networks, and Erdős-Rényi random graphs. This is a continuation of the hypothesis

made at the end of Chapter 4, that the correlation between adjacent subpopulations is largely driven by the local network structure. We show that the correlation between adjacent subpopulations is mostly determined by the network structure around the two focal subpopulations.

5.2 General metapopulation networks

In this chapter, we study the effect of metapopulation network structure in three network configurations: small metapopulation networks with $P = 4$ subpopulations, generalised star networks, and Erdős-Rényi random networks. In all network configurations, the underlying infectious disease model is the stochastic endemic infection model for a general metapopulation that we introduced in Chapter 4, Section 4.2. We assume throughout that all subpopulations are epidemiologically identical, that is, they are the same size and have the same epidemic parameters representing a measles-like endemic disease in the UK: $N = 10^5$, $R_0 = 17$, $\gamma^{-1} = 13$ days, $\mu = 5.5 \times 10^{-5}$ days^{-1} and $\epsilon = 5.5 \times 10^{-5}$ days^{-1}. For simplicity we also assume that the coupling between interacting subpopulations is fixed and equal to $\sigma = 0.1$.

In this section we define the three network configurations (Section 5.2.1) and introduce notation and concepts to describe properties of the metapopulation network structure (Section 5.2.2).

5.2.1 Network configurations

We introduce three network configurations that we use throughout this chapter to support analytic results: small metapopulation networks with $P = 4$ subpopulations, generalised star networks, and Erdős-Rényi random networks. When we refer to the focal subpopulations, we refer to the two subpopulations between which we are measuring the correlation.

Small networks ($P = 4$)

There are six connected metapopulation networks with $P = 4$ subpopulations, shown in Figure 5.1 and presented in order from least to most edges. Note that we have already studied two of these networks in Chapter 4: the star network (Figure 5.1a), and the complete network (Figure 5.1f).

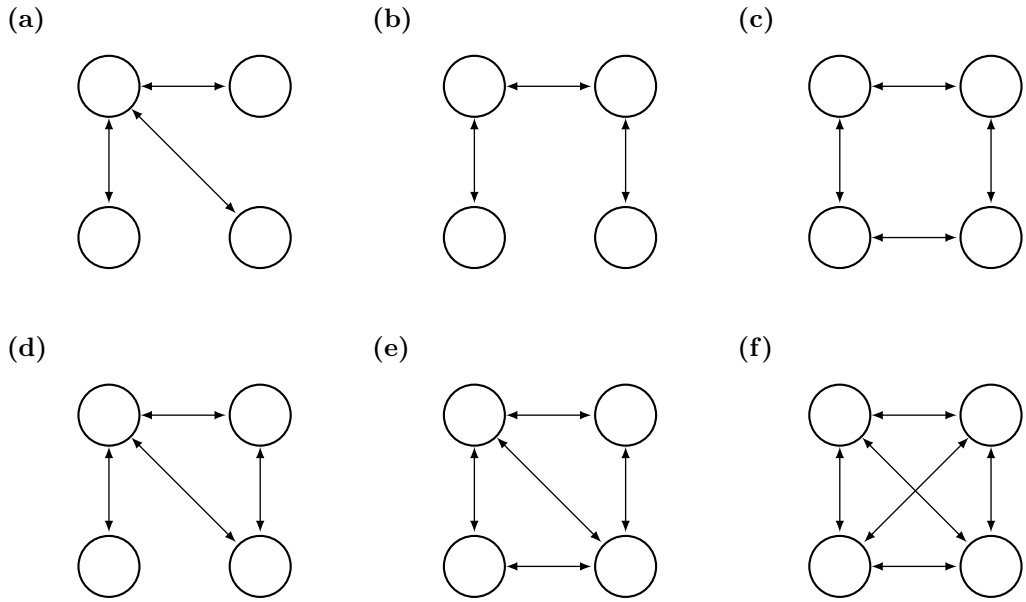


Figure 5.1. All connected networks with $P = 4$ subpopulations. Arrows (edges) between pairs of subpopulations show interaction between them. We assume throughout that all subpopulations are the same size, have the same epidemic parameters, and the coupling between all interacting pairs of subpopulations is $\sigma = 0.1$.

By studying such small networks we can exhaustively consider all network configurations; then, by comparing appropriate pairs of networks, we can observe the effect of adding in a single additional edge: for example, we can compare the correlations in the two networks shown in Figure 5.1b and 5.1c, which differ only by the addition of the bottom edge. We will use results in these small networks to motivate the study of more complex networks and to justify some assumptions that we make in Section 5.3.4.

Generalised star networks

The generalised star network can be fully defined by three parameters: $k_1 \in \mathbb{N}$, the number of subpopulations adjacent to focal subpopulation 1 only; $k_2 \in \mathbb{N}$, the number of subpopulations adjacent to focal subpopulation 2 only; and $k_3 \in \mathbb{N}$, the number of subpopulations adjacent to both focal subpopulations. This network configuration is illustrated in Figure 5.2.

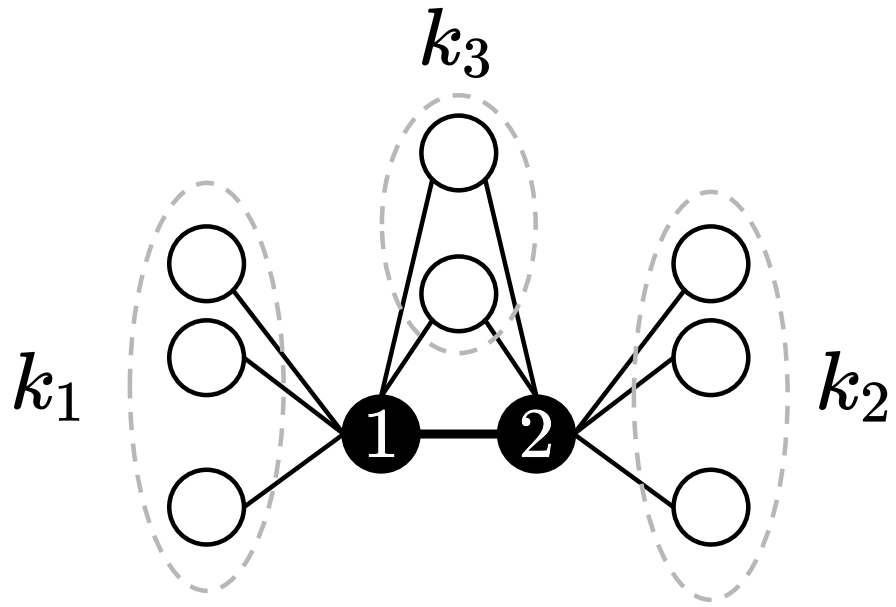


Figure 5.2. Structure of the generalised star network: there are $k_1 \in \mathbb{N}$ subpopulations adjacent to focal subpopulation 1 only, $k_2 \in \mathbb{N}$ subpopulations adjacent to focal subpopulation 2 only, and $k_3 \in \mathbb{N}$ subpopulations adjacent to both focal subpopulations. We assume throughout that all subpopulations are the same size, have the same epidemic parameters, and the coupling between all interacting pairs of subpopulations is $\sigma = 0.1$.

Erdős-Rényi random networks

Finally we consider random metapopulation networks generated using the Erdős-Rényi $G(P, q)$ model, which is defined as a metapopulation with P subpopulations where each pair of subpopulations interact (i.e. are coupled) with probability $q \in [0, 1]$, independent of other subpopulation pairs. In the literature this is usually written as $G(N, p)$, but we use alternative notation to avoid confusion with the subpopulation size, which we denote by N.

We choose the Erdős-Rényi network parameters P and q according to the following distributions: $P \sim \text{Uniform}\{10, 20\}$ and $q \sim \text{Uniform}(\log(P)/P, 2\log(P)/P)$. The range of P is chosen so that we can feasibly generate many network realisations and estimate the pairwise correlation between all subpopulations (this process will be slower on larger networks). The upper and lower limits of q are chosen such that network realisations are sparse (that is, the number of edges $|E|$ is not close to the maximum number of edges: $|E| \ll P(P-1)/2$), but also connected. If a realisation $G(P, q)$ is not connected then we generate new realisations (with the same P and q) until we generate a network that is connected. A realisation of an Erdős-Rényi $G(P, q)$ network ($P = 16, q = 0.32$) is shown in Figure 5.3.

5.2.2 Network definitions for metapopulation networks

To be able to describe the structure of general metapopulation networks, we introduce the following notation and definitions.

Let G be a metapopulation network with vertex set $V(G)$ and edge set $E(G)$, and let $x, y \in V(G)$ be the two focal subpopulations. We define the distance, $d \geq 1$, between x and y to be the length of the shortest path in the metapopulation from x to y. If $d = 1$ then there is an edge between x and y and we say that x and y are adjacent.

We define the neighbourhood of subpopulation x, which we denote $\mathcal{N}_x$, to be the set of subpopulations that incident to x in the network, that is, the set of subpopulations that interact directly with subpopulation x. We define the neighbourhood of the two focal subpopulations x and y, denoted $\mathcal{N}_{x\cup y}$, to be the set of subpopulations that interact with either x or y, excluding x and y themselves: $\mathcal{N}_{x\cup y} = (\mathcal{N}_x \cup \mathcal{N}_y)\backslash\{x, y\} \subset V(G)$. We define the common neighbourhood of the two focal subpopulations, denoted $\mathcal{N}_{x\cap y}$, to be the set of subpopulations that interact with both x and y: $\mathcal{N}_{x\cap y} = \mathcal{N}_x \cap \mathcal{N}_y \subset \mathcal{N}_{x\cup y}$. The density of a metapopulation network G is a measure of the density of edges and is

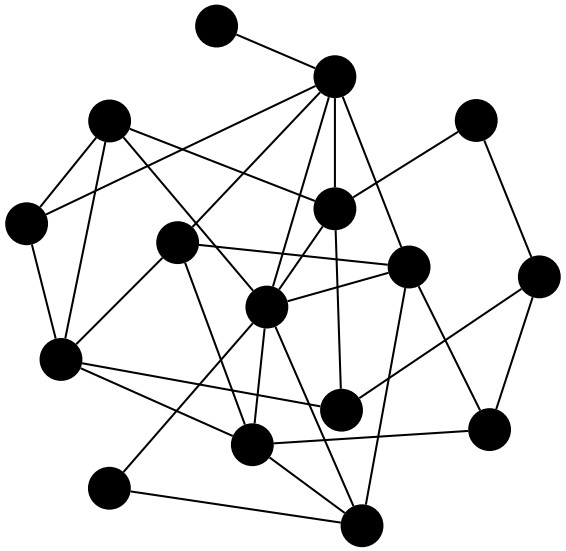


Figure 5.3. A single realisation of an Erdős-Rényi $G(P, q)$ random network with $P = 16$ and $q = 0.32$. We assume throughout that all subpopulations are the same size, have the same epidemic parameters, and the coupling between all interacting pairs of subpopulations is $\sigma = 0.1$.

calculated as $2|E|/(P(P-1))$, where $|E|$ is the number of edges in G.

Let $V' \subset V(G)$ be a subset of the subpopulations. The subgraph induced on V', denoted $G[V']$, is a subgraph of G whose vertex set is V' and whose edge set is the subset of edges $E' \subset E(G)$ with both ends are incident to vertices in V'. For the two focal subpopulations x and y, we denote the local network to be the subgraph induced on the set of subpopulation comprising x, y and their neighbourhood $\mathcal{N}_{x\cup y}$, that is, the subgraph $G[\mathcal{N}_{x\cup y} \cup \{x, y\}]$. When discussing the local network, we will sometimes refer to full metapopulation network G as the global network. Finally, we define the peripheral network of x and y to be the subgraph induced on the set of subpopulations not in their neighbourhood, that is, the subgraph $G[V(G)\backslash\mathcal{N}_{x\cup y}]$.

5.3 Estimating the correlation using a diffusion approximation

In this section we outline the method to numerically estimate the correlation between the prevalence of infection in two subpopulations within a general metapopulation network.

By approximating the $2P$-dimensional continuous-time Markov model of endemic disease dynamics (previously defined in Chapter 4, Section 4.2) by a diffusion process and making some additional simplifying assumptions (to follow, in Section 5.3.4), we can estimate the correlation between any pair of subpopulations in a general network. We support assumptions made in this section with the results of stochastic simulations on the small metapopulation networks with $P = 4$ subpopulations.

5.3.1 The Fokker-Planck approximation

Let $(\boldsymbol{X}(t), t \geq 0)$ be a $2P$-dimensional continuous-time Markov process describing endemic SIR infection for P interacting subpopulations of size N, as described in Chapter 4, Section 4.2. Under this definition, recall that $X_{2i-1}(t) = S_i(t)$ and $X_{2i}(t) = I_i(t), i = 1, \ldots, P$. Let $\boldsymbol{C} = (C_{ij})$ be the covariance matrix, where $C_{ij} = cov(X_i, X_j)$. Note that we are only concerned with covariances between the prevalence in different subpopulations, that is, of the form $cov(I_i, I_j), i \neq j$ (because $\boldsymbol{C}$ is symmetric, we can assume without loss of generality that $i < j$).

In the large-population limit ($N \to \infty$), we can approximate the discrete-state Markov process $(\boldsymbol{X}(t), t \geq 0)$ by a continuous-state diffusion process. Let $W(\boldsymbol{x}, \boldsymbol{r})$ be the transition rate from state $\boldsymbol{x}$ to state $\boldsymbol{x} + \boldsymbol{r}$. The multidimensional Fokker-Planck equation for the diffusion process is

$$\frac{\partial f}{\partial t} = -\sum_{i=1}^{P} \frac{\partial}{\partial x_i}(A_i f) + \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P} \frac{\partial^2}{\partial x_i \partial x_j}(B_{ij} f), \tag{5.1}$$

where $f = f(\boldsymbol{x}, t)$ is the probability density function of the diffusion process $(\boldsymbol{X}(t), t \geq 0)$, and

$$A_i(\boldsymbol{x}) = \sum_{\boldsymbol{r}} r_i W(\boldsymbol{x}, \boldsymbol{r}) \tag{5.2}$$

$$B_{ij}(\boldsymbol{x}) = \sum_{\boldsymbol{r}} r_i r_j W(\boldsymbol{x}, \boldsymbol{r}). \tag{5.3}$$

In Chapter 1 we described the behaviour of the first- and second-order moments of the diffusion process and showed that the covariance matrix $\boldsymbol{C}$ is the solution to the

Lyapunov equation

$$\sum_k a_{ik}C_{kj} + \sum_k a_{jk}C_{ki} = -b_{ij} \tag{5.4}$$

$$\iff \boldsymbol{aC} + \boldsymbol{Ca}^T = -\boldsymbol{b}, \tag{5.5}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are defined as

$$a_{ij} = \frac{\partial A_i}{\partial x_j}(\boldsymbol{x}^*) \tag{5.6}$$

$$b_{ij} = B_{ij}(\boldsymbol{x}^*), \tag{5.7}$$

and $\boldsymbol{x}^*$ is the endemic equilibrium state of the diffusion process.

5.3.2 Equivalence to multivariate normal moment closure approximation

We show that the linearised diffusion approximation is equivalent to the multivariate normal moment closure approximation that we used in Chapters 2 and 4, conditional on the structure of the constant matrix $\boldsymbol{b}$. Let $(\boldsymbol{X}(t), t \geq 0)$ be a $2P$-dimensional continuous-time Markov process whose rates are of the form $\alpha X_u X_v, \alpha \in \mathbb{R}$. We do not need to consider rates of the form αX_u, as both approximation methods only impact terms with non-linear rates; we do not consider third- or higher-order rates as there are no rates of this form in the stochastic endemic infection model. We demonstrate equivalence for first-order events (that is, transitions corresponding to the standard basis vectors: $\boldsymbol{e}_k, k = 1, \ldots, P$, where $\boldsymbol{e}_k$ is the vector with a 1 in the kth coordinate and 0s elsewhere), but the results hold more broadly. The rates of the process with first-order events are given by

$$W(\boldsymbol{X}, \boldsymbol{e}_k) = \sum_u \sum_v \alpha_{uv}^k X_u X_v, \quad k = 1, \ldots, P,$$

where $\alpha_{uv}^k \in \mathbb{R}$ are constants.

First we derive the ODE for the time evolution of $C_{ij} = cov(X_i, X_j)$ using a moment

closure approximation. The ODE for the time evolution of $\mathbb{E}[X_i X_j]$:

$$\begin{aligned}\frac{d\mathbb{E}[X_i X_j]}{dt} &= \mathbb{E}\left[(\sum_u \sum_v \alpha^i_{uv} X_u X_v) X_j + (\sum_u \sum_v \alpha^j_{uv} X_u X_v) X_i\right] \\ &= \sum_u \sum_v \alpha^i_{uv} \mathbb{E}[X_u X_v X_j] + \alpha^j_{uv} \mathbb{E}[X_u X_v X_i].\end{aligned}$$

Then, by making a second-order multivariate normal moment closure approximation (as described in Chapter 1, Section 1.2.2), the ODE for the time evolution of $C_{ij} = cov(X_i, X_j)$ is

$$\frac{dC_{ij}}{dt} = \sum_u \sum_v \alpha^i_{uv} X_u C_{vj} + \alpha^i_{uv} X_v C_{uj} + \alpha^j_{uv} X_u C_{vi} + \alpha^j_{uv} X_v C_{ui}.$$

At endemic equilibrium $dC_{ij}/dt = 0$ and so we get

$$0 = \sum_u \sum_v \alpha^i_{uv} X_u C_{vj} + \alpha^i_{uv} X_v C_{uj} + \alpha^j_{uv} X_u C_{vi} + \alpha^j_{uv} X_v C_{ui}. \tag{5.8}$$

We can also derive Equation (5.8) using a diffusion approximation. By linearising the Fokker-Planck equation for the diffusion process approximating $(\boldsymbol{X}(t), t \geq 0)$ around the endemic equilibrium $\boldsymbol{X}^*$, $a_{ij} = (\partial A_i / \partial x_j)(\boldsymbol{x}^*)$ is given by

$$\begin{aligned}a_{ij} &= \frac{\partial A_i}{\partial X_j}(\boldsymbol{X}^*) \\ &= \frac{\partial}{\partial X_j}\left(\sum_u \sum_v \alpha^i_{uv} X_u X_v\right)\Bigg|_{\boldsymbol{X}=\boldsymbol{X}^*} \\ &= \sum_u \alpha^i_{uj} X^*_u + \sum_v \alpha^i_{jv} X^*_v.\end{aligned}$$

Substituting this into the Lyapunov equation $-b_{ij} = \sum_k a_{ik} C^*_{kj} + \sum_k a_{jk} C^*_{ki}$, we get

$$\begin{aligned}-b_{ij} &= \sum_k \left(\sum_v \alpha^i_{kv} X^*_v + \sum_u \alpha^i_{uk} X^*_u\right) C^*_{kj} + \sum_k \left(\sum_v \alpha^j_{kv} X^*_v + \sum_u \alpha^j_{uk} X^*_u\right) C^*_{ki} \\ &= \sum_u \sum_v \alpha^i_{uv} X^*_v C^*_{uj} + \alpha^i_{uv} X^*_u C^*_{vj} + \alpha^j_{uv} X^*_v C^*_{ui} + \alpha^i_{uv} X^*_u C^*_{vi}.\end{aligned}$$

For first-order events, then $b_{ij} = 0, \forall i, j$ and so this is the same result that we derived using the multivariate normal moment closure approximation (Equation (5.8)). For more complex events (e.g. infection) this may or may not be true, so we must include

the additional condition that $b_{ij} = 0$; we show in Section 5.3.3 that this holds for our stochastic process and certain values of i and j.

We have shown that the diffusion approximation described above is equivalent to the multivariate normal moment closure approximation, up to $b_{ij} = 0$. This means that estimates of the correlation obtained using the diffusion approximation are comparable to estimates of the correlation that have previously obtained using the multivariate normal moment closure approximation. In the remainder of this section we explain how we solve the Lyapunov equation (Equation (5.4)) for the correlation between prevalence of infection in any pair of subpopulations.

5.3.3 The structure of Lyapunov equation for endemic disease dynamics

For a general stochastic process, the solution to the Lyapunov equation (Equation (5.4)) may be non-trivial to compute. However, for our model of endemic SIR infection in a general metapopulation network, we can exploit the structure of the Jacobian $\boldsymbol{a}$, the constant matrix $\boldsymbol{b}$, and the covariance matrix $\boldsymbol{C}$ to allow us to find the solution for each $\rho_{ij}, i < j$, without solving the full Lyapunov equation.

Structure of the Jacobian, $\boldsymbol{a}$

Recall that $\boldsymbol{a}$ is defined as

$$a_{ij} = \frac{\partial A_i}{\partial x_j}(\boldsymbol{x}^*),$$

where $A_i(\boldsymbol{x}) = \sum_{\boldsymbol{r}} r_i W(\boldsymbol{x}, \boldsymbol{r})$. We show that $\boldsymbol{a}$ is a sparse matrix, that is, $a_{ij} = 0$ for $i \in \{2k-1, 2k\}$ and $j = 2k' - 1, k' \neq k$. For an example visualisation of the structure $\boldsymbol{a}$, see Figure 5.4a.

First we note that $X_{2k-1} = S_k$ and $X_{2k} = I_k$, and so, through the infection event, A_{2k-1} and A_{2k} both contain terms with $X_{2k-1} = S_k$ and some or all of $X_{2j} = I_j, j = 1, \ldots, P$, depending on the metapopulation network structure. However, neither A_{2k-1} nor A_{2k} contain terms with $X_{2k'-1} = S_{k'}, k' \neq k$, and therefore $(\partial A_{2k-1}/\partial x_{2k'-1})(\boldsymbol{x}^*) = 0$ and $(\partial A_{2k}/\partial x_{2k'-1})(\boldsymbol{x}^*) = 0$.

This is useful as it reduces the number of terms on the left-hand side of the Lyapunov equation (Equation (5.4)).

Structure of the constant matrix, $\boldsymbol{b}$

Recall that $\boldsymbol{b}$ is defined as

$$b_{ij} = B_{ij}(\boldsymbol{x}^*) = \sum_{\boldsymbol{r}} r_i r_j W(\boldsymbol{x}^*, \boldsymbol{r}).$$

We can show that $\boldsymbol{b}$ is a block diagonal matrix where off-diagonal elements are 0. For an example visualisation of the structure $\boldsymbol{b}$, see Figure 5.4b.

This result follows from the fact that $W(\boldsymbol{x}, \boldsymbol{r})$ is only non-zero for a very small set of $\boldsymbol{r}$. First, by the definition of the endemic disease dynamics, $r_i \in \{-1, 0, 1\}, \forall i$. Second, as there is no explicit movement between subpopulations, then change in subpopulation i means that there is no change in subpopulation $j \neq i$. So $\boldsymbol{r}$ is a sparse vector where the only non-zero elements are of the form $(r_{2k-1}, r_{2k}) \in \{(-1, +1), (-1, 0), (1, -1), (1, 0)\}$, representing the events of infection, recovery, death of an infected individual, and death of a recovered individual, respectively, in subpopulation k. Therefore, $b_{ij} \neq 0$ only when $j = 2k$ and $i = j - 1, j$.

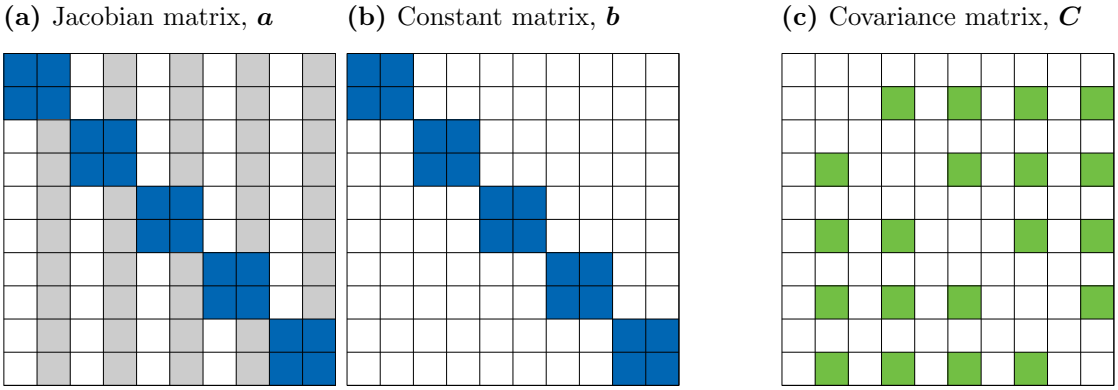
We showed in Section 5.3.2 that if $b_{ij} = 0$ then the diffusion approximation and the moment closure approximation are equivalent. Moreover, if $b_{ij} = 0$ then we may not need to calculate the matrix $\boldsymbol{b}$ at all to solve the Lyapunov equation (Equation (5.4)).

Structure of the covariance matrix, $\boldsymbol{C}$

In a metapopulation network with P subpopulations, $\boldsymbol{C} \in \mathbb{R}^{2P \times 2P}$. However, as we only want to find the correlations between prevalence in pairs of subpopulations, we only need to find expressions for $cov(I_i, I_j), i < j$, that is, for $C_{2i,2j} = cov(X_{2i}, X_{2j}), i < j$. In a metapopulation with P subpopulations, there are at most $P(P-1)/2$ covariances of this form. Therefore, instead of solving the full Lyapunov equation, we can simply solve a system of $P(P-1)/2$ carefully chosen simultaneous equations. For an example visualisation of the covariances of interest in $\boldsymbol{C}$, see Figure 5.4c.

5.3.4 Solving the Lyapunov equation for prevalence covariances

We now describe how we use the Lyapunov equation (Equation (5.4)) to find expressions for $\rho_{ij} = corr(I_i, I_j), i < j$, in terms of the coupling σ. In a metapopulation with P subpopulations, there are at most $P(P-1)/2$ correlations of the form $\rho_{ij}, i < j$: due



(a) Jacobian matrix, $\boldsymbol{a}$ **(b)** Constant matrix, $\boldsymbol{b}$ **(c)** Covariance matrix, $\boldsymbol{C}$

Figure 5.4. Example visualisation of the structure of the matrices in the Lyapunov equation $\boldsymbol{aC} + \boldsymbol{Ca}^T = -\boldsymbol{b}$ for a metapopulation network with $P = 5$ subpopulations. **(a)** Structure of the Jacobian matrix $\boldsymbol{a} \in \mathbb{R}^{2P \times 2P}$, defined by $a_{ij} = \partial A_i / \partial x_j(\boldsymbol{x}^*)$. Blue squares denote strictly non-zero elements and white squares denote zero elements; grey squares denote elements that may be zero, depending on the metapopulation network structure. **(b)** Structure of the constant matrix $\boldsymbol{b} \in \mathbb{R}^{2P \times 2P}$, defined by $b_{ij} = \sum_{\boldsymbol{r}} r_i r_j W(\boldsymbol{x}^*, \boldsymbol{r})$. Blue squares denote strictly non-zero elements and white squares denote zero elements. **(c)** Structure of the covariance matrix $\boldsymbol{C} \in \mathbb{R}^{2P \times 2P}$, defined by $C_{ij} = cov(X_i, X_j)$. Green squares denote covariances of interest, that is, covariances of the form $cov(X_{2i}, X_{2j}) = cov(I_i, I_j), i \neq j)$.

to symmetries in the metapopulation network structure, some of these correlations may be the same. In summary, our approach is to take a subset of the equations of the form in Equation (5.4), then simplify them so that they are in terms of $\rho_{ij}, i < j$, only, then solve for ρ_{ij}.

Consider the following system of $P(P-1)/2$ equations from the Lyapunov equation:

$$\sum_{k=1}^{2P} a_{i'k}C_{kj'} + \sum_{k=1}^{2P} a_{j'k}C_{ki'} = -b_{i'j'}$$

where $i' = 2i$ and $j' = 2j, i < j$. By the arguments outlined in Section 5.3.3, $b_{i'j'} = 0$, so we have

$$\sum_{k=1}^{2P} a_{i'k}C_{kj'} + \sum_{k=1}^{2P} a_{j'k}C_{ki'} = 0. \tag{5.9}$$

The left-hand side of the $i'j'$-th Lyapunov equation (Equation (5.9)) contains three types of covariance terms: variances in infection prevalence, of the form $C_{i'i'} = Var(I_i)$; covariances between infection prevalence in different subpopulations, of the form $C_{i'k'} = cov(I_i, I_k), k \neq i$; and between-population covariances of the form $C_{k'-1,i'} = cov(S_k, I_i), k \neq i$ (the coefficient of terms for covariances of the form $C_{i'-1,i'} = cov(S_i, I_i)$ is always zero, by the arguments outlined in Section 5.3.3). We make two simplifying assumptions so that the left-hand side contains only covariance terms for the correlations $\rho_{ij}, i < j$:

$$\sum_{k \neq j} a_{i,2k}\rho_{jk} + \sum_{k \neq i} a_{j,2k}\rho_{ik} = -(a_{ii} + a_{jj}),$$

which we then solve to find the correlations. The two assumptions are outlined below and the effect of these assumptions is explored numerically in Section 5.3.5.

Assumption 1: Variance in prevalence is the same in all subpopulations

The first simplifying assumption that we make is that the variance in prevalence is the same in all subpopulations, which we will denote by V. If we divide Equation (5.9) through by V then the left-hand side of the $i'j'$-th equation now contains only two types of terms: correlations between prevalence in different subpopulations, $\rho_{ij}, i < j$; and terms of the form $C_{k'-1,i'}/V$.

Assumption 2: $\hat{C}_{S_iJ_j}$ terms are negligibly small

The second simplifying assumption that we make is that terms of the form $a_{j',k'-1}C_{k'-1,i'}/V$ are negligibly small and can therefore be ignored; in earlier notation, this is of the form $a\hat{C}_{SI}/C_{II}$. In Chapters 2 and 4 we make the same claim about terms of this form when using the moment closure approximation, and therefore extend it here (without further justification) to general metapopulation networks. By ignoring the effect of these terms, the left-hand side of Equation (5.9) now contains $\rho_{ij}, i < j$ terms and constant terms only.

5.3.5 Comparison to stochastic simulations

We evaluate the effect of the two assumptions by comparing our estimate of the correlation using the Fokker-Planck equation to the results of stochastic simulations.

First, we compare the variance in infection prevalence in each of the subpopulations in the small metapopulation networks with $P = 4$ subpopulations (Figure 5.5). Even in these small networks it is clear that the variances are not all equal; in fact, we only have equality when all of the subpopulations are epidemiologically and topologically identical (for example, network 6, which is the complete network).

However, the combination of both assumptions has only a small effect on the correlation. We compare our estimate of the correlation using the diffusion approximation to the results of stochastic simulations in small metapopulation networks where $P = 4$ (Figure 5.6). This shows that we consistently overestimate the correlation between subpopulations, but that the magnitude of this difference is small (mean 0.041, maximum 0.058), and is comparable to the difference between the multivariate normal correlation and our approximation for two subpopulations as shown in Chapter 2.

5.4 The effect of metapopulation network structure on the correlation

Using the diffusion approximation method, we can easily calculate the correlation between prevalence in any pair of subpopulations in a metapopulation network, without needing to derive any ODEs. We use this method to explore the effect of metapopulation network structure on the correlation between subpopulations.

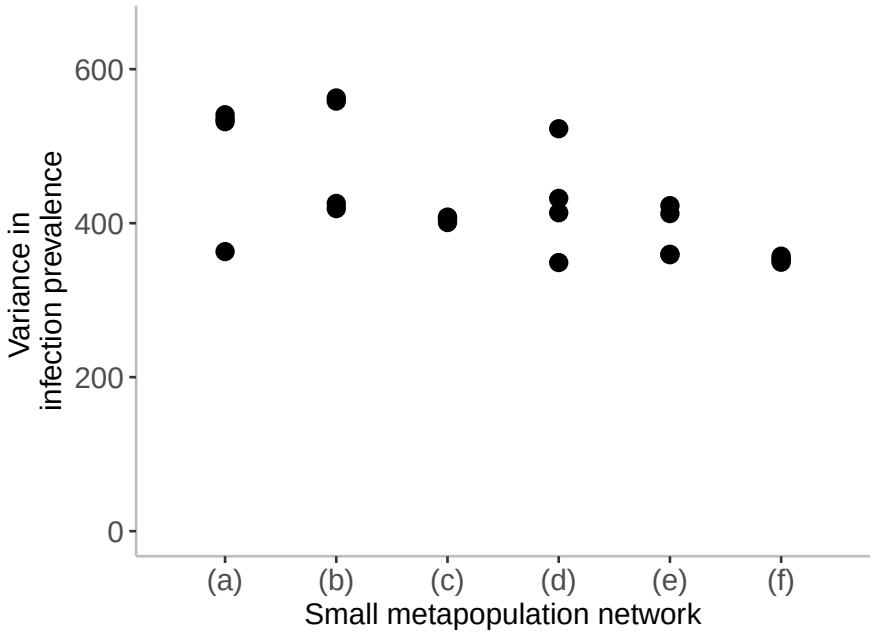


Figure 5.5. Comparing the variance in infection prevalence in each of the six small metapopulation networks with $P = 4$ subpopulations (labelled (a)-(f); network configurations are shown in Figure 5.1). For each network configuration, the variance in infection prevalence is calculated in each of the four subpopulations by simulating the stochastic endemic infection model. Points show the mean variance of the prevalence in each subpopulation in each network, where the mean is taken over 1000 realisations of the stochastic process.

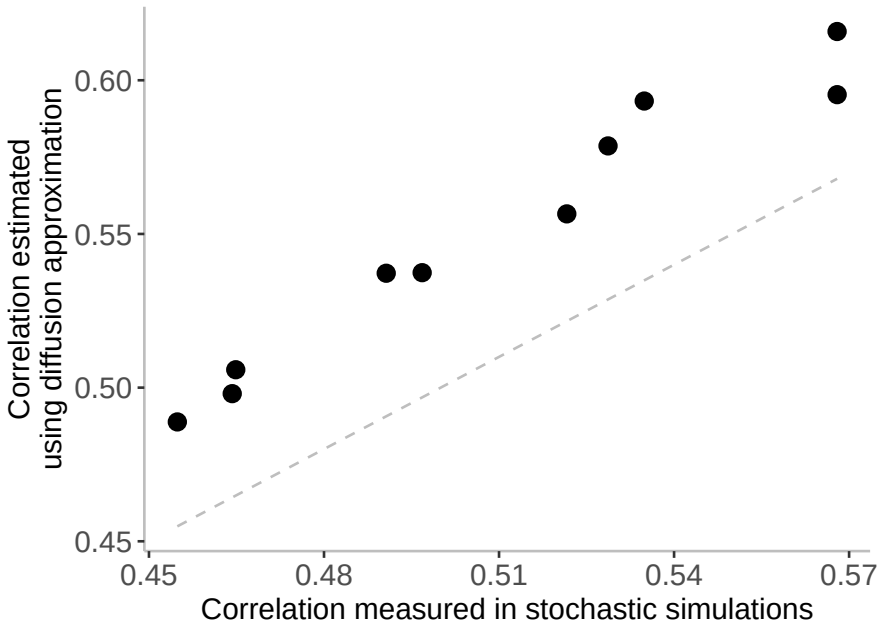


Figure 5.6. Comparing the correlation estimated using the diffusion approximation to the results of stochastic simulations in small metapopulation networks with $P = 4$ subpopulations. For the simulations, we generate 1000 realisations of the stochastic endemic infection model and measure the correlation between each pair of subpopulations in each of the six metapopulation networks; points show the mean correlation over the 1000 realisations for each subpopulation pair. Grey dashed line shows $y = x$, for reference.

First, we show that the correlation between subpopulations decays exponentially with the network distance between them, which motivates us to restrict our attention to adjacent subpopulations only. For each of the three network configurations (small metapopulations with $P = 4$ subpopulations, generalised star networks, and Erdős-Rényi random network), we consider the effect of local network properties (neighbourhood size, common neighbourhood size and neighbourhood density, as defined in Section 5.2.2) on the correlation between a pair of adjacent subpopulations. For Erdős-Rényi random networks we also show that peripheral network structure (also defined in Section 5.2.2) has only a small effect on the correlation. Finally, we use a simple multiple linear regression model to predict the correlation between adjacent subpopulations from the local network properties.

5.4.1 Distance between subpopulations

In Chapter 4 we showed that the correlation between subpopulations in the k-regular tree network decays exponentially with the distance distance d between them: the correlation between two subpopulations distance d apart is given by

$$\rho_d = \left(\frac{k\sigma + \xi - \sqrt{\sigma^2 k^2 - (2\xi\sigma - 4\sigma^2)k + 4\sigma^2 + \xi^2}}{2(k-1)\sigma} \right)^d, \tag{5.10}$$

where $\sigma \in [0, 1]$ is the coupling between interacting subpopulations, and $\xi \approx \xi' = \epsilon/(\mu(R_0 - 1))$. In this section we show that this general relationship holds for more general network structures.

First we consider small metapopulation networks with $P = 4$ subpopulations. As the distance d between the subpopulations increases, the correlation between the subpopulations decreases (Figure 5.7). However, with so few data points, it is difficult to make any general conclusions about the functional form of the relationship between the two variables in general metapopulation networks.

In Erdős-Rényi random networks this relationship is more clear. The correlation decays exponentially with the distance between the subpopulations (Figure 5.8). There is some overlap between the correlation at different distances, which is caused by other aspects of the network structure, which we discuss in detail shortly.

The distance between subpopulations is clearly an important factor in determining the correlation: as the distance increases then the correlation quickly decays to zero,

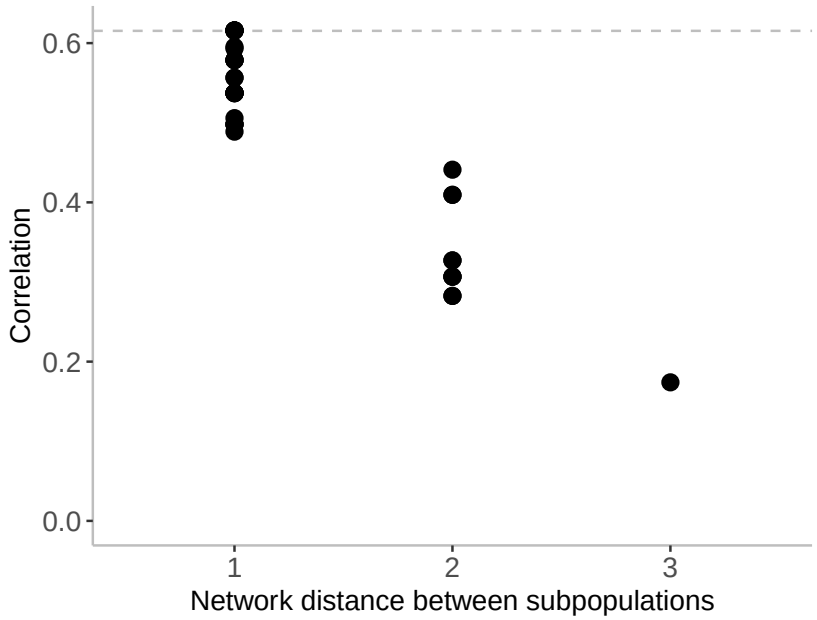


Figure 5.7. Effect of network distance on the correlation between infection prevalence in small metapopulation networks with $P = 4$. The correlation between each pair of subpopulations in each of the six small metapopulation networks is estimated using the diffusion approximation method described in Section 5.3. Grey dashed line indicates the correlation between any pair of subpopulations in the complete network for $\sigma = 0.1$, for reference.

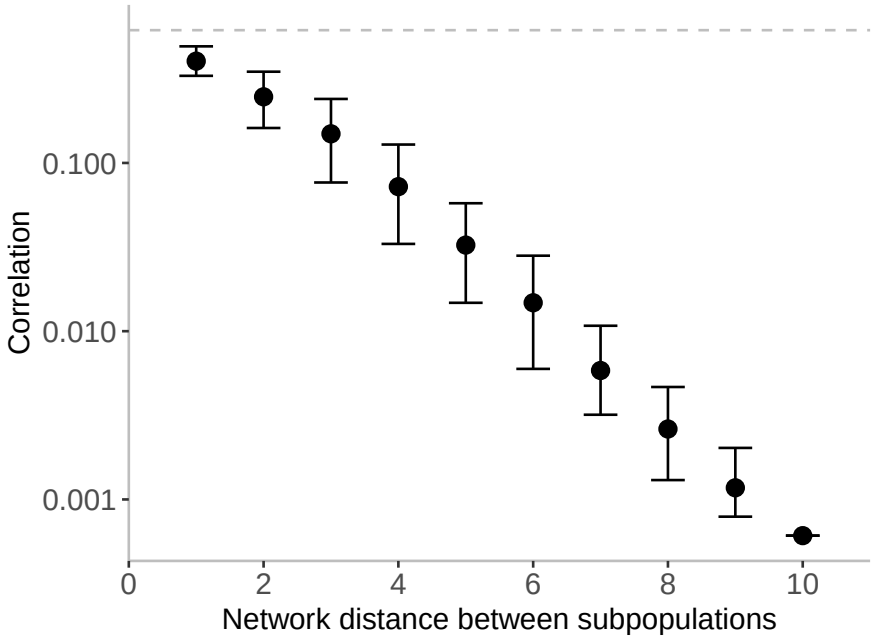


Figure 5.8. Effect of network distance on the correlation between infection prevalence in Erdős-Rényi random networks. The correlation between each pair of subpopulations in 1000 network realisations is estimated using the diffusion approximation method described in Section 5.3. Points show mean correlation for each distance and errorbars show 2.5-th and 97.5-th percentiles; note the log scale on the y-axis. Grey dashed line indicates the correlation between any pair of subpopulations in the complete network for $\sigma = 0.1$, for reference.

regardless of the rest of the network structure. When the distance d between subpopulations is small, there is still considerable variation in the observed correlation. For the remainder of this section we aim to determine the drivers of the correlation between prevalence of infection in adjacent subpopulations in general metapopulation networks. For each of the three network configurations defined in Section 5.2 (small metapopulation networks with $P = 4$ subpopulations, generalised star networks, and Erdős-Rényi random networks), we consider the effect of neighbourhood size, common neighbourhood size and neighbourhood density on the correlation between adjacent subpopulations (that is, where $d = 1$).

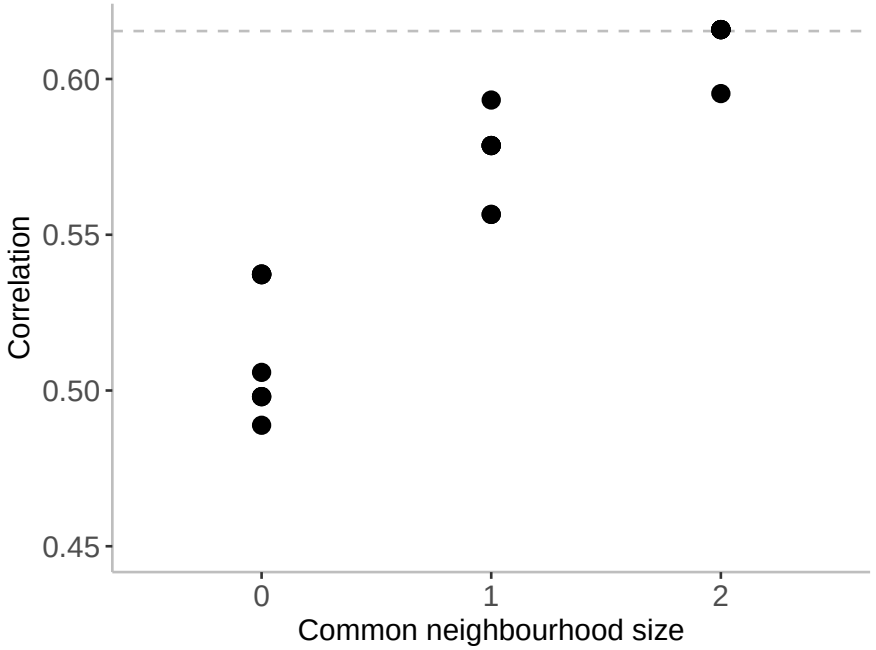


Figure 5.9. Effect of common neighbourhood size on the correlation between infection prevalence in adjacent subpopulations in small metapopulation networks with $P = 4$ subpopulations. The correlation between each pair of subpopulations in each of the six small metapopulation networks is estimated using the diffusion approximation method described in Section 5.3. Grey dashed line indicates the correlation between any pair of subpopulations in the complete network for $\sigma = 0.1$, for reference.

5.4.2 Local network structure in small networks ($P = 4$)

First we consider the effect of common neighbourhood size and neighbourhood density on the correlation between prevalence of infection in two adjacent subpopulations in small metapopulation networks with $P = 4$ subpopulations. We do not discuss the effect of neighbourhood size because for almost all pairs of adjacent subpopulations in the small metapopulation networks the neighbourhood size is 2.

Effect of common neighbourhood size

The size of the common neighbourhood in the small metapopulation networks with $P = 4$ subpopulation ranges from 0 to 2. As the common neighbourhood size increases then the correlation between adjacent subpopulations also increases (Figure 5.9).

Effect of neighbourhood density

Recall that the neighbourhood density is only defined when the neighbourhood size is greater than or equal to 2. In the small metapopulation networks with $P = 4$ subpopulations, the maximum neighbourhood size is 2, so the neighbourhood density can only take two values: 1, if there is an edge between the two neighbours, or 0, if there is no edge. Increasing the neighbourhood density from 0 to 1 effectively introduces a 4-cycle into the network.

Using color to highlight appropriate pairs of networks, we see that as neighbourhood density increases, the correlation also increases (Figure 5.10). The magnitude of the increase is largest (equal to 0.048) when the common neighbourhood size is zero (subpopulation pair 2), as in this network there are no 3-cycles. In comparison, when the common neighbourhood size is greater than or equal to 1, the magnitude of the increase is smaller (0.008, 0.022 and 0.02 for subpopulation pairs 1, 3 and 4, respectively). This is because there is already at least one 3-cycle present in these networks.

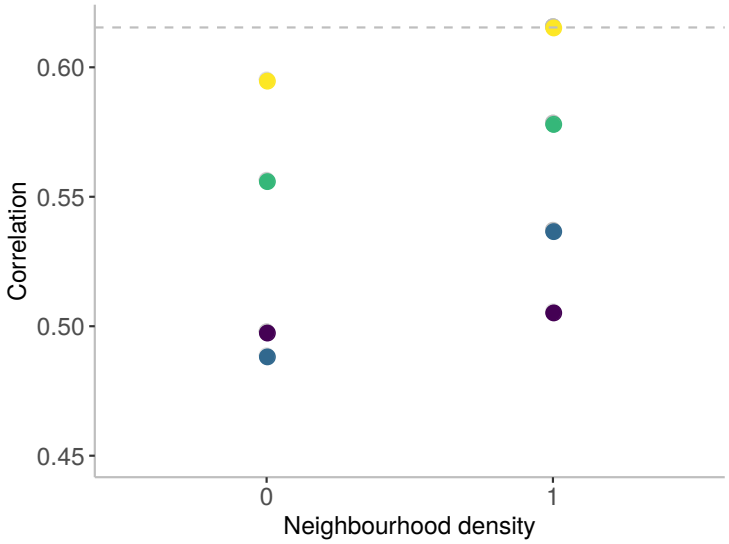
5.4.3 Local network structure in generalised star networks

The structure of the generalised star network is fully defined by the parameters k_1, k_2 (the number of subpopulations adjacent only to focal subpopulations 1 and 2, respectively) and k_3 (the number of subpopulations incident to both focal subpopulations). We can define the local network properties in terms of these parameters: the neighbourhood size of the two focal subpopulations is $k_1 + k_2 + k_3$ and the common neighbourhood size is k_3. By definition, there is no interaction between any of the neighbours, so the neighbourhood density is always zero. In this section, we consider the effect of neighbourhood size and common neighbourhood size on the correlation between prevalence of infection in the two focal subpopulations, where $k_1, k_2, k_3 \in [0, 5]$.

Effect of neighbourhood size

As the neighbourhood size increases, then the correlation between the two focal subpopulations decreases (Figure 5.11). Moreover, for a given neighbourhood size n, the correlation increases as the common neighbourhood size increases. Therefore, the correlation between the two focal subpopulations in the generalised star network with neighbourhood size n is bounded below by the correlation for a common neighbourhood size of zero ($k_3 = 0$), and above by the correlation for $k_3 = n$.

(a)



(b)

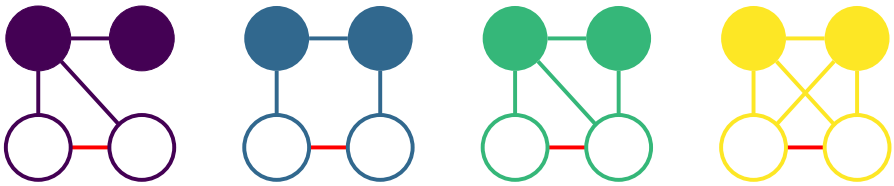


Figure 5.10. Effect of neighbourhood density on the correlation between infection prevalence in adjacent subpopulations in small metapopulation networks with $P = 4$ subpopulations. **(a)** The correlation between each pair of subpopulations in each of the six small metapopulation networks is estimated using the diffusion approximation method described in Section 5.3. Colours of points show pairs of networks that differ only in the neighbourhood density (i.e. neighbourhood size and common neighbourhood size are the same). Grey dashed line indicates the correlation between any pair of subpopulations in the complete network for $\sigma = 0.1$, for reference. **(b)** Visualisation of the four network pairs, where the additional edge (which increases the neighbourhood density from 0 to 1) is shown in red. Correlation is measured between the filled nodes.

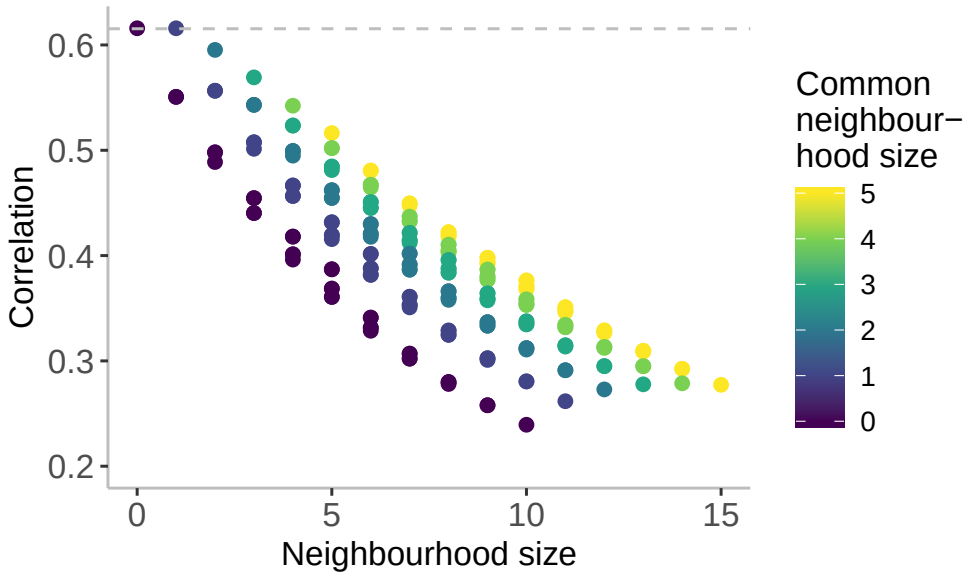


Figure 5.11. Effect of neighbourhood size ($k_1 + k_2 + k_3$) on the correlation between infection prevalence in the two focal subpopulations in the generalised star network, where $k_1, k_2, k_3 \in \{0, 1, \ldots, 5\}$. The correlation between the focal subpopulations for each network configuration is estimated using the diffusion approximation method described in Section 5.3. Colours of points shows the common neighbourhood size (k_3). Grey dashed line indicates the correlation between any pair of subpopulations in the complete network for $\sigma = 0.1$, for reference.

For a given neighbourhood size and common neighbourhood size, there is still some variation in the correlation between the focal subpopulations, which we show arises as a result of the distribution of neighbours between the focal subpopulations. We define a measure of unbalance, u, to describe how evenly the neighbours are distributed between the two focal subpopulations: $u = -1 + 2\max(k_1, k_2)/(k_1 + k_2 + k_3) \in [0, 1]$. A value of 0 indicates that the neighbours are evenly distributed between the two subpopulations; a value of 1 indicates that the neighbours are all adjacent to one focal subpopulation only.

The distribution of neighbours in the generalised star network has a small, but noticeable, effect. For a fixed neighbourhood size and common neighbourhood size, the correlation is highest in neighbourhoods where the neighbours are evenly distributed between the two focal subpopulations. For example, when $k_3 = 0$ (no common neighbours, Figure 5.12) and $k_1 + k_2 = n$, the correlation is lowest in the network where all n neighbours are adjacent to one focal subpopulation, and highest in the network where each focal subpopulation has $n/2$ neighbours. The magnitude of the effect on the correlation is very small: for the values of k_1, k_2 and k_3 that we consider, the maximum effect that this network structure has on the correlation is 0.026 (when $k_1 + k_2 = 5, k_3 = 0$).

Effect of common neighbourhood size

In the generalised star network the common neighbourhood size has little effect on the correlation between adjacent subpopulations (Figure 5.13a). That the maximum correlation is decreasing and the minimum correlation is increasing is an artifact of the network structure and that $k_1, k_2.k_3 \in [0, 5]$ only. On the other hand, as the relative common neighbourhood size increases, then the correlation also increases (Figure 5.13b). We note that neighbourhood size still plays an important role: for a fixed relative common neighbourhood size, the correlation is higher in smaller neighbourhoods.

5.4.4 Local network structure in Erdős-Rényi random networks

We now consider how neighbourhood size, common neighbourhood size and neighbourhood density affect the correlation between infection prevalence in two adjacent subpopulations in Erdős-Rényi random networks. We also show that peripheral network structure has only a small effect on the correlation.

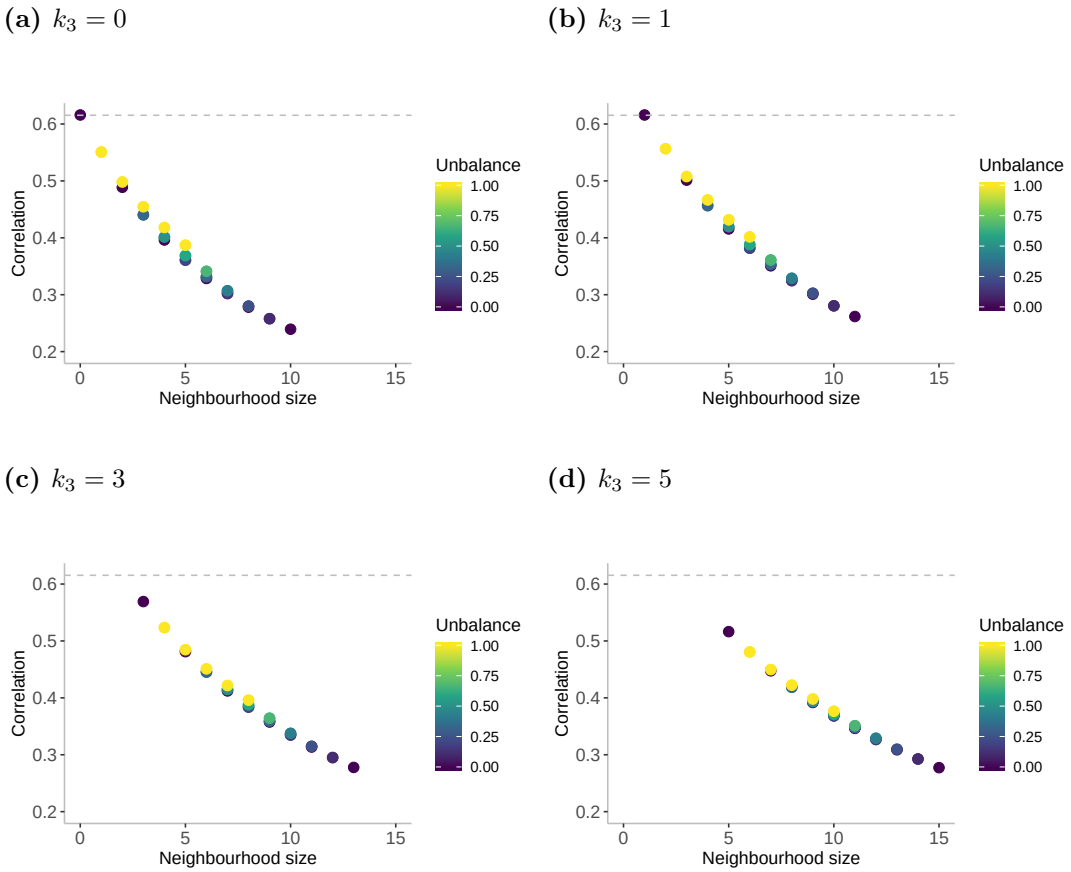


Figure 5.12. Effect of distribution of neighbours on the correlation between infection prevalence in the two focal subpopulations in the generalised star network for **(a)** $k_3 = 0$, **(b)** $k_3 = 1$, **(c)** $k_3 = 3$ and **(d)** $k_3 = 5$ ($k_1, k_2 \in \{0, 1, \ldots, 5\}$ throughout). The correlation between the focal subpopulations for each network configuration is estimated using the diffusion approximation method described in Section 5.3. Colours of points show the unbalance measure. The unbalance measure, u, describe how evenly the neighbours are distributed between the two focal subpopulations and is calculated as $u = -1 + 2\max(k_1, k_2)/(k_1 + k_2 + k_3) \in [0, 1]$ ($u = 0$ when the neighbours are evenly distributed between the two subpopulations, and $u = 1$ when the neighbours are all adjacent to one focal subpopulation only). Grey dashed line indicates the correlation between any pair of subpopulations in the complete network for $\sigma = 0.1$, for reference.

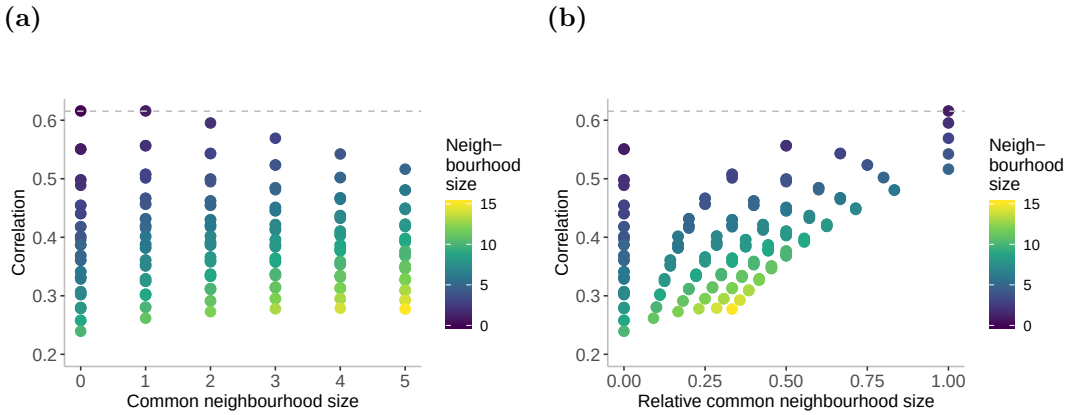


Figure 5.13. Effect of **(a)** common neighbourhood size and **(b)** relative common neighbourhood size on the correlation between infection prevalence in the two focal subpopulations in the generalised star network, where $k_1, k_2, k_3 \in \{0, 1, \ldots, 5\}$. The common neighbourhood size is given by $k_3 \in \{0, 1, \ldots, 5\}$ and the relative common neighbourhood size is calculated as $k_3/(k_1 + k_2 + k_3) \in [0, 1]$. The correlation between the focal subpopulations for each network configuration is estimated using the diffusion approximation method described in Section 5.3. Colours of points shows the neighbourhood size ($k_1 + k_2 + k_3 \in \{0, 1, \ldots, 15\}$). Grey dashed line indicates the correlation between any pair of subpopulations in the complete network for $\sigma = 0.1$, for reference.

Effect of peripheral network structure

In Chapter 4 we hypothesised that the correlation between adjacent subpopulations is mostly determined by the local network structure, that is, by the structure of the subgraph induced on the set of vertices comprising the focal subpopulations and their neighbours. We provide evidence for this assertion by comparing the correlation between adjacent subpopulations in the full metapopulation network (which we will refer to as the global correlation) to the correlation between the same pair of subpopulations in their local network (which we will refer to as the local correlation).

The global correlation is positively correlated with the local correlation (Pearson's correlation coefficient $r = 0.94$). For the majority of the network realisations, the local correlation is higher than the global correlation (Figure 5.14a), although the mean absolute difference between the two correlation measures is small (mean 0.031, maximum 0.089) (Figure 5.14b). Moreover, the correlation in the local network is a significant predictor of the correlation in the full network ($\beta = 0.8, p < 0.0001; R^2 = 0.876$). We therefore conclude that it is adequate to estimate the correlation in the local network, rather than the full network. An advantage of this is that it is quicker to estimate the correlation in the local network than the global network, as the local network is smaller.

In the rest of this section we look at the effect of neighbourhood size, common neighbourhood size and neighbourhood density on the local correlation; however, the conclusions that we make also hold for the global correlation.

Effect of neighbourhood size

The distribution neighbourhood size in the Erdős-Rényi random networks is similar to the generalised star networks (mean 5.99, standard deviation 2.05; Figure 5.15a). As with the generalised star networks, we observe that as the neighbourhood size increases, the local correlation between adjacent subpopulations decreases (Pearson correlation coefficient $r = -0.52$; Figure 5.15), and for a fixed neighbourhood size n, the local correlation increases as the common neighbourhood size increases. We also find that neighbourhood size is a significant predictor of the correlation in Erdős-Rényi random networks ($\beta = -0.011, p < 0.0001$).

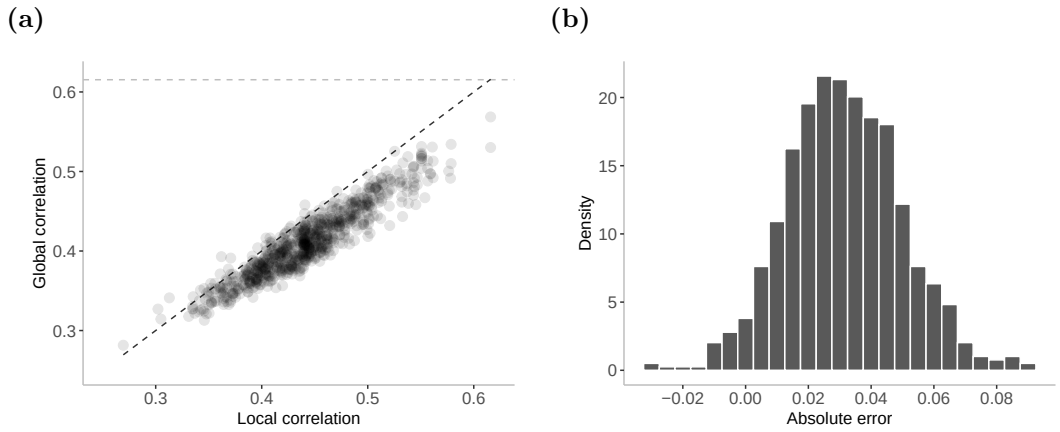


Figure 5.14. Comparing the local and global correlation between infection prevalence in adjacent subpopulations in Erdős-Rényi $G(P, q)$ random networks, where $P \sim \text{Uniform}\{10, 20\}$ and $q \sim \text{Uniform}(\log(P)/P, 2\log(P)/P)$. **(a)** For 1000 network realisations, the local and global correlation are estimated using the diffusion approximation method described in Section 5.3; the global correlation is estimated on the full network, whilst the local correlation is estimated on the subgraph induced on the set of vertices comprising the two focal subpopulations and their neighbours. Grey dashed line indicates the correlation between any pair of subpopulations in the complete network for $\sigma = 0.1$, and black dashed line indicates $y = x$, for reference. **(b)** The absolute error is calculated as the absolute difference between the two correlation measures.

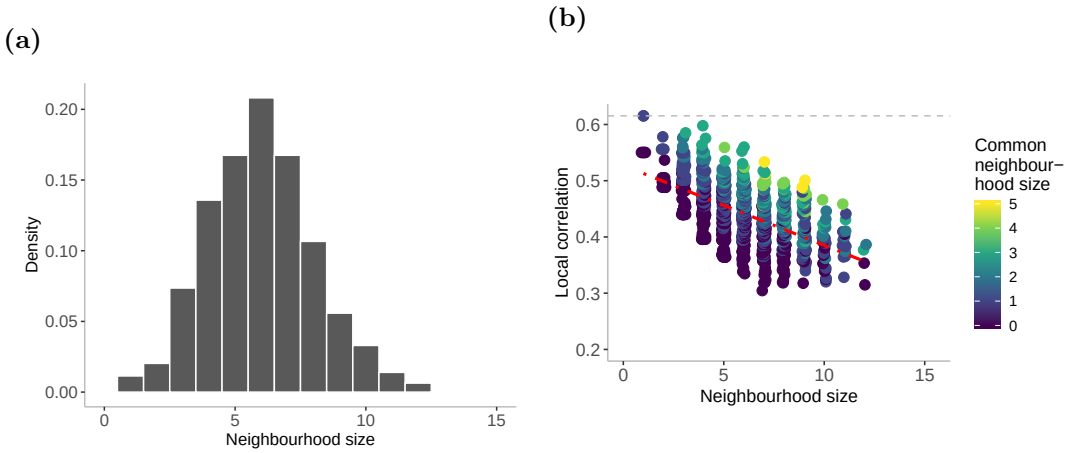


Figure 5.15. **(a)** Distribution of neighbourhood size for two randomly-chosen adjacent subpopulations in 1000 realisations of Erdős-Rényi $G(P, q)$ random networks, where $P \sim \text{Uniform}\{10, 20\}$ and $q \sim \text{Uniform}(\log(P)/P, 2\log(P)/P)$. **(b)** Effect of neighbourhood size on the correlation between infection prevalence in adjacent subpopulations in the same 1000 Erdős-Rényi random networks. The local correlation between subpopulations is estimated using the diffusion approximation method described in Section 5.3, on the subgraph induced on the set of vertices comprising the two focal subpopulations and their neighbours. Colour of points shows the common neighbourhood size; random horizontal jitter has been added to the points to improve readability, but has no meaning. Red dot-dashed line shows the simple linear regression for this relationship ($\beta = -0.011, p < 0.0001$). Grey dashed line indicates the correlation between any pair of subpopulations in the complete network for $\sigma = 0.1$, for reference.

Effect of common neighbourhood size

The common neighbourhood size in the Erdős-Rényi random networks ranges from 0 to 4 (Figure 5.16a). The correlation between adjacent subpopulations is more strongly correlated with the relative common neighbourhood size (Pearson's correlation coefficient $r = 0.63, p < 0.0001$; Figure 5.16b) than the common neighbourhood size ($r = 0.38, p < 0.0001$; not shown). Both neighbourhood size and relative common neighbourhood size are significant predictors of the correlation in Erdős-Rényi random networks, but relative common neighbourhood size has a better model fit ($\beta = 0.184, p < 0.0001$; $R^2 = 0.394$). There is still a considerable amount of variation in the correlation for a given relative common neighbourhood size, especially for common neighbourhood size equal to zero. As in the generalised star network, for a given relative common neighbourhood size, the correlation is higher in networks with smaller neighbourhoods.

Effect of neighbourhood density

The mean neighbourhood density in the Erdős-Rényi random networks is 0.27 (standard deviation 0.17) (Figure 5.17a). There are very few network realisations with neighbourhood density greater than 0.5 due to the distributions for the number of subpopulations P and the probability of interaction q that we use to generate the random networks. The neighbourhood density is positively correlated with the correlation (Pearson's correlation coefficient $r = 0.46$; Figure 5.17b), and is a significant predictor of the correlation in Erdős-Rényi random networks ($\beta = 0.117, p < 0.0001$).

5.4.5 Predicting the correlation between adjacent subpopulations

Neighbourhood size, common neighbourhood size (or relative common neighbourhood size) and neighbourhood density are all significant predictors of the correlation between prevalence of infection in adjacent subpopulations in Erdős-Rényi random networks. In this section we show that the local correlation (the correlation estimated in the local network) and the global correlation (the correlation estimated in the full metapopulation network) can be predicted from the local network structure, with high accuracy.

We use a multiple linear regression model to predict the local and global correlation between adjacent subpopulations in the Erdős-Rényi random networks. In both models we use the three local network properties (neighbourhood size, common neighbourhood size and neighbourhood density) as the independent variables. We also consider models

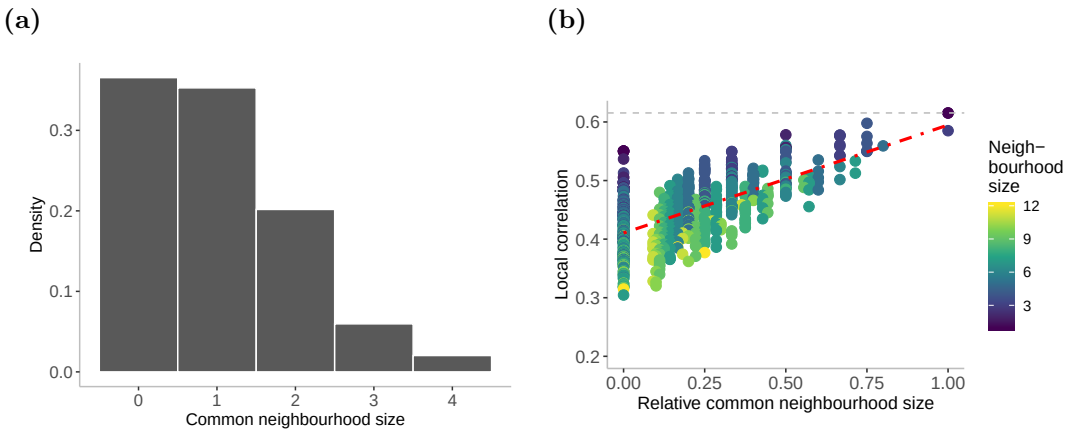


Figure 5.16. **(a)** Distribution of common neighbourhood size for two randomly-chosen adjacent subpopulations in 1000 realisations of Erdős-Rényi $G(P, q)$ random networks, where $P \sim \text{Uniform}\{10, 20\}$ and $q \sim \text{Uniform}(\log(P)/P, 2\log(P)/P)$. **(b)** Effect of relative common neighbourhood size (relative to neighbourhood size) on the correlation between infection prevalence in adjacent subpopulations in the same 1000 Erdős-Rényi random networks. The local correlation between subpopulations is estimated using the diffusion approximation method described in Section 5.3, on the subgraph induced on the set of vertices comprising the two focal subpopulations and their neighbours. Colour of points shows the neighbourhood size. Red dot-dashed line shows the simple linear regression for this relationship ($\beta = 0.147, p < 0.0001$). Grey dashed line indicates the correlation between any pair of subpopulations in the complete network for $\sigma = 0.1$, for reference.

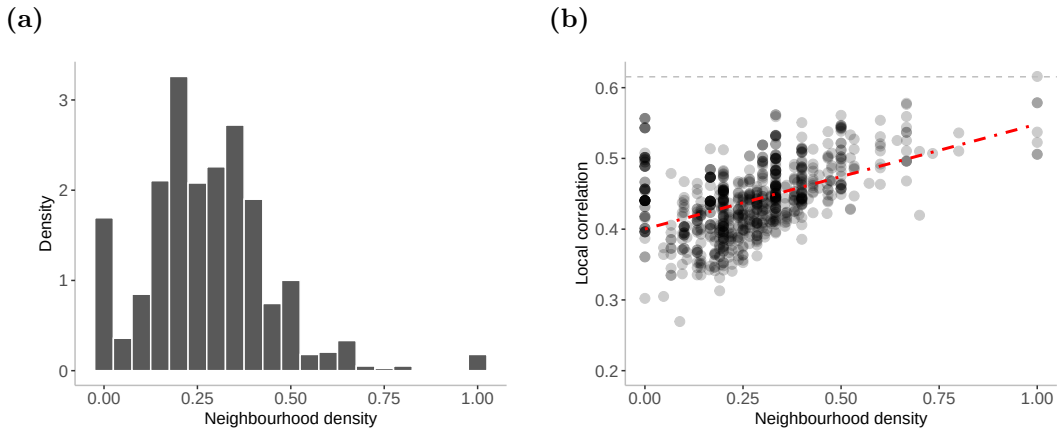


Figure 5.17. **(a)** Distribution of neighbourhood density for two randomly-chosen adjacent subpopulations in 1000 realisations of Erdős-Rényi $G(P, q)$ random networks, where $P \sim \text{Uniform}\{10, 20\}$ and $q \sim \text{Uniform}(\log(P)/P, 2\log(P)/P)$. **(b)** Effect of neighbourhood density on the correlation between infection prevalence in adjacent subpopulations in the same 1000 Erdős-Rényi random networks. The local correlation between subpopulations is estimated using the diffusion approximation method described in Section 5.3, on the subgraph induced on the set of vertices comprising the two focal subpopulations and their neighbours. Transparency of points is used to show point density. Red dot-dashed line shows the simple linear regression for this relationship ($\beta = 0.117, p < 0.0001$). Grey dashed line indicates the correlation between any pair of subpopulations in the complete network for $\sigma = 0.1$, for reference.

using relative common neighbourhood size, rather than common neighbourhood size, but it had very little effect on the results. For the model predicting the global correlation, we also include the total network size as a fourth independent variable. The structure of each of the regression models is summarised in Table 5.1.

Predicting the correlation in the local network

First, we aim to predict the local correlation. Neighbourhood size ($\beta_1 = -0.018, p < 0.0001$), common neighbourhood size ($\beta_2 = 0.03, p < 0.0001$) and neighbourhood density ($\beta_3 = 0.128, p < 0.0001$) are all significant predictors of the local correlation, with a model fit of $R^2 = 0.847$. We compare the predicted and actual local correlation (Figure 5.18). The model predictions do not show any bias, although there are a few outlying points where the difference between the actual and predicted correlation is surprisingly large.

The results of this model clearly shows the conflicting effect of the neighbourhood size and the common neighbourhood size. The net result of adding a new common neighbour is to increase the correlation: the correlation increases as a result of increasing the common neighbourhood size by 1, but also decreases as a result of increasing the neighbourhood size by 1.

Predicting the correlation in the global network

In Section 5.4.4 we showed that the correlation in the local network is a significant predictor of the correlation in the full network ($\beta = 0.8, p < 0.0001$; $R^2 = 0.876$). We show that we can predict the global correlation from (mostly) local network properties.

Again, neighbourhood size, common neighbourhood size and neighbourhood density are all significant predictors of correlation in the local network (β coefficients are given in Table 5.1, Model 3), and the overall model fit is $R^2 = 0.75$. If network size (which is not a local network property, but is a very basic measure of the global network structure) is included as an additional variable, the model fit improves to $R^2 = 0.836$, which is comparable to the fit of the model predicting the local correlation. Again, comparing the predicted and actual global correlation shows that our predictions are unbiased (Figure 5.19).

	Model structure	β_0	β_1	β_2	β_3	β_4	R^2
1	local correlation = β_0 + β_1n'hood size + $\beta_2$common n'hood size + β_3n'hood density	0.484	-0.018	0.03	0.128	-	0.847
2	local correlation = β_0 + β_1n'hood size + $\beta_2$relative common n'hood size + β_3n'hood density	0.456	-0.013	0.167	0.122	-	0.856
3	global correlation = β_0 + β_1n'hood size + $\beta_2$common n'hood size + β_3n'hood density	0.442	-0.014	0.025	0.1	-	0.75
4	global correlation = β_0 + β_1n'hood size + $\beta_2$relative common n'hood size + β_3n'hood density	0.42	-0.01	0.137	0.096	-	0.744
5	global correlation = β_0 + β_1n'hood size + $\beta_2$common n'hood size + β_3n'hood density + $\beta_4$network size	0.496	-0.011	0.023	0.086	-0.004	0.836
6	global correlation = β_0 + β_1n'hood size + $\beta_2$relative common n'hood size + β_3n'hood density + $\beta_4$network size	0.475	-0.007	0.124	0.082	-0.004	0.83

Table 5.1. Structure of multiple linear regression models used to predict the local and global correlation between adjacent subpopulations in general metapopulation networks using local network properties (models 1-4), and local network properties and network size (models 5 and 6). All variables included are significant predictors of the correlation ($p < 0.0001$). We have abbreviated neighbourhood to n'hood, for brevity.

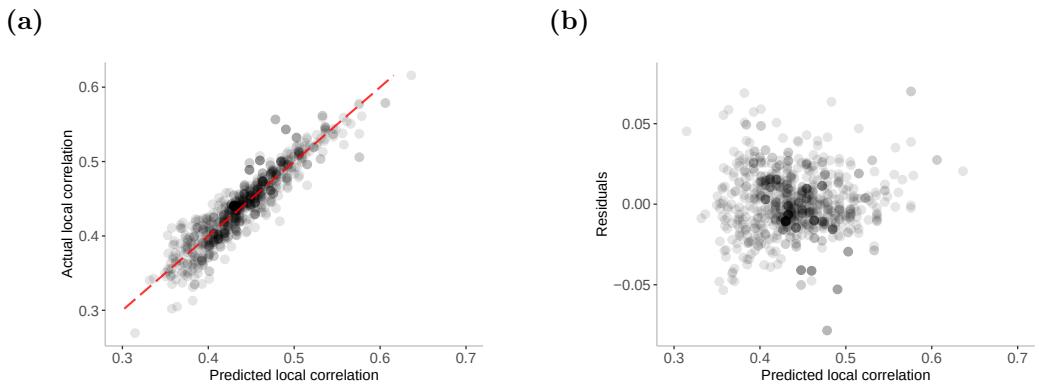


Figure 5.18. **(a)** Comparing the actual and predicted local correlation between adjacent subpopulations in 1000 realisations of Erdős-Rényi $G(P, q)$ random networks, where $P \sim \text{Uniform}\{10, 20\}$ and $q \sim \text{Uniform}(\log(P)/P, 2\log(P)/P)$. The actual local correlation between subpopulations is estimated using the diffusion approximation method described in Section 5.3, on the subgraph induced on the set of vertices comprising the two focal subpopulations and their neighbours. The predicted correlation is obtained from the multiple linear regression model with neighbourhood size, common neighbourhood size and neighbourhood density as the independent variables (Model 1, Table 5.1), with a model fit of $R^2 = 0.847$. Red dashed line shows $y = x$, for reference. **(b)** Residual plot of the regression model. In both plots, transparency of points is used to show point density.

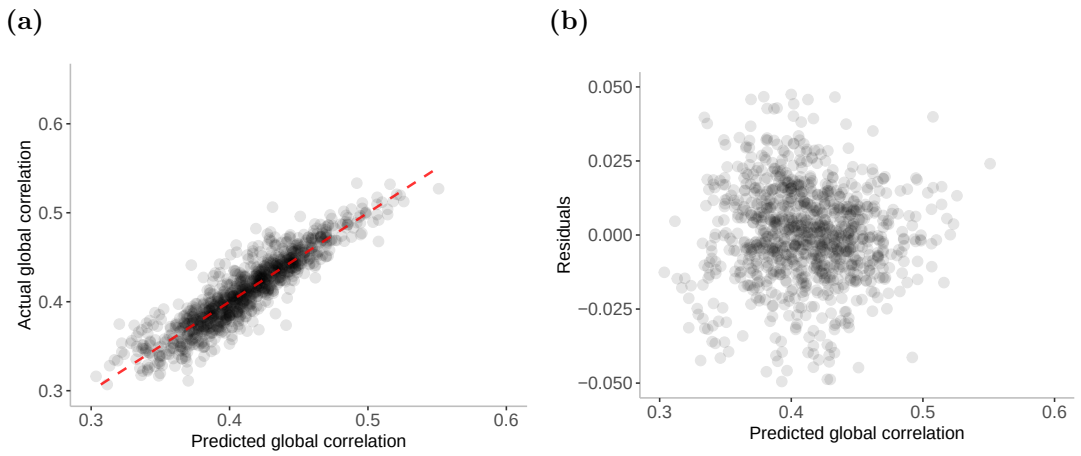


Figure 5.19. **(a)** Comparing the actual and predicted global correlation between adjacent subpopulations in 1000 realisations of Erdős-Rényi $G(P, q)$ random networks, where $P \sim \text{Uniform}\{10, 20\}$ and $q \sim \text{Uniform}(\log(P)/P, 2\log(P)/P)$. The actual global correlation between subpopulations is estimated using the diffusion approximation method described in Section 5.3, on the full network. The predicted global correlation is obtained from the multiple linear regression model with neighbourhood size, common neighbourhood size, neighbourhood density and network size as the independent variables (Model 5, Table 5.1, with a model fit of $R^2 = 0.836$. Red dashed line shows $y = x$, for reference. **(b)** Residual plot of the regression model. In both plots, transparency of points is used to show point density.

5.5 Discussion

In this chapter we use present a method to estimate the correlation between prevalence of infection in two subpopulations in a general metapopulation network. By approximating the continuous-time Markov model of endemic disease dynamics by a diffusion process and making some additional reasonable simplifying assumptions, we can calculate the correlation between any pair of subpopulations in a general network.

This builds and improves on the work from Chapter 4 by proposing a simpler method by which the correlation between any pair of subpopulations in a general metapopulation network can be estimated. Using the Fokker-Planck approximation we can easily numerically generate the set of simultaneous equations to solve for the correlations between subpopulations with little effort. In contrast, the moment closure approximation method required us to derive ODEs for the time evolution of each of the covariances $cov(I_i, I_j), i < j$.

Although this method is certainly useful, it still has some limitations. First, to ensure that we can solve the system of simultaneous equations for the prevalence correlations, we assume that the variance in infection prevalence is the same in all subpopulations. This assumption is where the two estimates of the correlation (using the moment closure approximation and Fokker-Planck approximation) differ. Although we showed that this assumption is not true, we have also shown that the difference between the two estimates is small and so a reasonable price to pay for generality. The second limitation is that unlike the small or symmetric networks we considered in Chapters 2 and 4, we cannot obtain any meaningful analytic expressions to relate the correlation, coupling and network structure; instead, all our analyses are numerical. The final limitation is on the size of the metapopulation network. The size of the system of simultaneous equations that we need to solve grows quadratically with the number of subpopulations, so as the number of subpopulations increases, this method quickly becomes computationally-intensive. In this chapter we have only considered metapopulation networks where $P \in [10, 20]$; larger networks will obviously take longer to solve, but we do not quantify this here.

The other contribution of this chapter is to use the diffusion approximation method to explore the effect of the metapopulation network structure on the correlation between subpopulations. This addresses the hypothesis made in Chapter 4, that the correlation between adjacent subpopulations is largely determined by the local network structure. We consider the effect of neighbourhood size, common neighbourhood size and neigh-

bourhood density.

Results from both the generalised star networks (Section 5.4.3) and the Erdős-Rényi random networks (Section 5.4.4) show that as the neighbourhood size increases then the correlation between adjacent subpopulations decreases. Furthermore, neighbourhood size is a significant predictor of both the local and global correlation in Erdős-Rényi random networks ($\beta = -0.018, -0.011$, respectively).

On the other hand, as the common neighbourhood size in the small metapopulation networks (Section 5.4.2) and the Erdős-Rényi random networks increases, the correlation between adjacent subpopulations increases. In the generalised star network there was no relationship between the common neighbourhood size and the correlation, as a result of the way that the network is defined and the limited range of values that were used for the parameters k_1, k_2 and k_3. However, for a given common neighbourhood size n, we can put bounds on the correlation in the generalised star network: the correlation is bounded above by the correlation in the network where $k_3 = n$ and $k_1 = k_2 = 0$, and bounded below by 0 (for $k_3 = 0$ and $k_1 + k_2 \to \infty$).

When comparing networks of different sizes (generalised star network and Erdős-Rényi random networks), the relative common neighbourhood size is more informative than the common neighbourhood size. For example, for a regression model with a single independent variable, relative common neighbourhood size is a better predictor of the correlation than common neighbourhood size in the Erdős-Rényi random networks ($R^2 = 0.394$, compared to $R^2 = 0.152$). In the generalised star network only relative common neighbourhood size is a significant predictor of the correlation ($\beta = 0.167$, $R^2 = 0.274$). The fit of regression models with either common neighbourhood size or relative common neighbourhood size is comparable (for the local correlation $R^2 = 0.847, 0.856$, respectively; for the global correlation, including network size as an independent variable, $R^2 = 0.836, 0.83$, respectively), but interpreting the results of the model using common neighbourhood size is more intuitive.

Interaction between neighbours, measured by neighbourhood density, also acts to increase the correlation between adjacent subpopulations. Results from the small metapopulation networks show this result is most significant when the common neighbourhood size is zero, since increasing the neighbourhood density then introduces 4-cycles into the networks. In the Erdős-Rényi random networks, neighbourhood density is a significant predictor of both the local and global correlation ($\beta = 0.127, 0.086$, respectively).

Using only three simple local network properties (neighbourhood size, common neigh-

bourhood size and neighbourhood density) we are able to predict the local and global correlation ($R^2 = 0.847, 0.75$, respectively). The predictive ability of both models could be improved by including more complex local network properties, such as neighbourhood size of each focal subpopulations and edge-density of the common and uncommon neighbourhoods. We have also shown that the predictive ability of the global correlation can be improved by including the size of the full metapopulation network (which increases the coefficient of determination from $R^2 = 0.75$ to $R^2 = 0.836$); further improvements could be made by including additional information about the structure of the peripheral network. However, we note that this would probably only offer small improvements.

From the results in Chapter 4, we know that for a fixed neighbourhood size n, the correlation (in the local and global network) is bounded above by the correlation between any pair of subpopulations in the complete network, which is approximately equal to $\sigma/(\xi' + \sigma), \xi' = \epsilon/(\mu(R_0 - 1))$. We hypothesise that the local correlation is bounded below by correlation in the generalised star network where $k_1 = \lfloor n/2 \rfloor, k_2 = n - k_1$ and $k_3 = 0$; this makes intuitive sense, since both the common neighbourhood size and the neighbourhood density are zero. However, the lower bound of the global correlation is less clear and needs further study; our hypothesis is that the lower bound is equal to zero, in a network that looks locally like the generalised star network with $k_1 = \lfloor n/2 \rfloor, k_2 = n - k_1$ and $k_3 = 0$, but where each of the neighbours have infinitely many neighbours.

We believe that this analysis is a useful contribution towards understanding the complex relationship between network structure and metapopulation endemic disease dynamics, but there is still more that can be done. First, we have only considered Erdős-Rényi random networks $G(P, q)$ where $P \sim \text{dUniform}(10, 20)$ and $q \sim \text{Uniform}(\log(P)/P, 2\log(P)/P)$; using these distributions, we do not generate realisations of networks that are more well-connected (e.g. with larger neighbourhood sizes, relative common neighbourhood sizes, or high neighbourhood density). We also have not considered other random network models (e.g. the Barabasi-Albert model or Watts-Strogatz model); such models may more accurately describe real-world metapopulation networks, or have desired properties (such as a heterogeneous degree distribution, clustering, or small-world properties). Nonetheless, we believe that these additional analyses would not significantly affect our main conclusions.

5.6 Conclusions

It is difficult to estimate the correlation between infection prevalence in two subpopulations in a general metapopulation using the moment closure approximation method described in Chapter 2 and 4. In this chapter we outline an alternative, but largely equivalent, method for estimating the correlation that is simpler and more easily generalisable than the moment closure method. By approximating the Markov model with a diffusion approximation and making some additional reasonable simplifying assumptions, we can calculate the correlation between any pair of subpopulations in a general network. We use this method to explore the effect of network structure on the correlation between adjacent subpopulations. Our results show that the correlation between adjacent subpopulations is largely determined by the local network structure, and are a useful contribution towards a fuller understanding the effect of network structure on endemic disease dynamics.

Chapter 6

Conclusions and further work

In this thesis we explore the dynamics of endemic infection in a metapopulation network. Our initial aim was to explore the analytic relationship between the correlation between prevalence of infection and the coupling, σ, between two subpopulations. We used approximation methods to derive an analytic approximation for the correlation between prevalence of infection in a simple two-subpopulation network, and then in more complex and general networks. We used these methods to explore the effect of metapopulation network structure on the correlation between the prevalence of infection in pairs of subpopulations. We also considered practical aspects of using the analytic results to infer the coupling, which is often unknown, from the correlation between subpopulations.

The underlying model used throughout this thesis describes the dynamics of endemic infectious disease in a metapopulation, that is, where the population is divided into interacting subpopulations. The underlying disease dynamics represent a simple SIR dynamics: in almost all numerical simulations we consider a measles-like disease, although acknowledge that a realistic model of measles would include both seasonality and age-structure. The way that the interaction between subpopulations is defined is sufficiently general that it can describe multiple forms of heterogeneity, including spatial-, age- and risk-structured mixing.

In **Chapter 2** we use a multivariate moment closure approximation to derive a simple analytic relationship between the correlation between prevalence of infection and the coupling between two identical subpopulations. A particular strength of this relationship is that it can be fully defined using only the epidemic parameters. Moreover, we show that this approximation holds for a wide range of parameter values, including many

endemic childhood diseases. We also highlight that this relationship between the coupling and the correlation could be used to address the challenge of inferring the coupling between interacting subpopulations, especially in the absence of contact or mobility data. In **Chapter 3** we develop this idea further and consider practical elements of this approach. We show that a shorter observation period leads to increased variability in the estimated coupling, and a bias to overestimate the coupling when the true coupling is very small. Lower frequency observations and using recovery incidence data have little effect on the observed correlation and the subsequent estimated coupling when considered alone. However, when both limitations are applied together, this leads to significant overestimates of the coupling for very high coupling values, although we acknowledge that such high coupling values are likely unrealistic in real-world systems.

The rest of the thesis build on the results in Chapter 2 by considering correlations between subpopulations in general metapopulation networks. In **Chapter 4** we use the same multivariate normal moment closure approximation to derive analytic expressions for the correlation between subpopulations in symmetric metapopulation networks, specifically the complete network, the k-regular tree network and the star network. Of particular interest is that the correlation between any pair of subpopulations in the complete network is independent of the network size, even though network size had a marked effect on the correlation between adjacent subpopulations in both the k-regular tree network and the star network. This led us to hypothesise that the correlation between two adjacent subpopulations is driven by local network properties, specifically the number of neighbours and common neighbours that the two focal subpopulations have, plus any interactions between them. However, this chapter also demonstrated the challenges of estimating the correlation in general metapopulation networks using the moment closure approximation, since many ODE equations for the first- and second-order moments need to be derived.

We address this challenge in **Chapter 5**: instead of the moment closure approximation we approximate the Markov process by a diffusion process, which allows us to numerically estimate the correlation between any pair of subpopulations in a general metapopulation network. Notably, this method is equivalent to the multivariate moment closure approximation, up to an additional assumption about the variance in prevalence of infection. The advantage of this method is that it is much simpler to implement and generalise. In the second half of Chapter 5 we use this method to explore the hypothesis set out at the end of Chapter 4. We show that the correlation between adjacent subpopulations is largely determined by the local network structure. Moreover, the correlation

can be predicted from local network properties, namely neighbourhood size, common neighbourhood size and neighbourhood density.

We envision that future work following this thesis comprises two related directions. The first is to further generalise the diffusion approximation method for inferring the correlation between subpopulations. We can extend the underlying epidemiological model by including additional complexity, such as age-structure, seasonality, and additional infection compartments, all of which will help to better describe the underlying infectious disease dynamics. Whilst we anticipate that we would still be able to use the diffusion approximation method, there may be some new challenges. New compartments (introduced either through new infection compartments or age-structured mixing) will give rise to new covariances; to be able to use the same approach as Chapter 5, we will either need to make additional assumptions about these covariances, or increase the size of the system of simultaneous equations that we need to solve.

As outlined at the end of Chapter 5, we can also perform additional analyses to ensure that our results hold for all network configurations, both for a broader definition or Erdos-Renyi networks and for other random network configurations. This extension may more accurately describe real-world metapopulation networks, or give rise to networks with desired properties (such as a heterogeneous degree distribution, clustering, or small-world properties). We anticipate, however, that these extensions will not significantly affect the results presented here, since we have clearly demonstrated that it is local network properties that have the greatest effect on the correlation between adjacent subpopulations.

We can further generalise the underlying metapopulation network structure by considering heterogeneity in epidemic parameters, subpopulation size or coupling between subpopulations. Our assumption about the variance in prevalence of infection may no longer hold, so this work will therefore also need to consider how these variances scale with such changes.

The second direction of future work is to further develop our method for inferring the coupling between subpopulations from the observed correlation. We can consider other limitations to the observation process, such as the effect of underreporting or unobserved cases, or observing incidence of infection. We also aim to understand the interaction between different data limitations, such as observing aggregated recovery incidence. Beyond this, we can consider how these results change in a general metapopulation network, as well as how the method might be impacted by missing observed

correlation data between some pairs of subpopulations.

In summary, this thesis makes useful contributions towards understanding the dynamics of endemic infection in a metapopulation network, as well as taking steps towards a method for estimating the strength of interaction between subpopulations from observable data.

Appendix A

Appendix to Chapter 2

A.1 Derivation of the ODE system for stochastic endemic infection model for two populations

For the coupled stochastic epidemic model, we can approximate the stochastic process by the following system of eight ODEs. The ODEs for the five first-order central moments are:

$$\frac{d\bar{S}}{dt} = \mu N - \frac{\beta}{N}(1-\sigma)(C_{SI} + \bar{S}\bar{I}) - \frac{\beta}{N}\sigma(\hat{C}_{SI} + \bar{S}\bar{I}) - \epsilon\bar{S} - \mu\bar{S} \tag{A.1}$$

$$\frac{d\bar{I}}{dt} = \frac{\beta}{N}(1-\sigma)(C_{SI} + \bar{S}\bar{I}) + \frac{\beta}{N}\sigma(\hat{C}_{SI} + \bar{S}\bar{I}) + \epsilon\bar{S} - \gamma\bar{I} - \mu\bar{I} \tag{A.2}$$

$$\begin{aligned}\frac{dC_{SS}}{dt} &= \mu N + \frac{\beta}{N}\bar{S}\bar{I} + \epsilon\bar{S} - \mu\bar{S} - 2\left(\frac{\beta}{N}\bar{I} + \epsilon + \mu\right)C_{SS} - \frac{\beta}{N}(1-\sigma)(2\bar{S}-1)C_{SI} \\ &\quad - \frac{\beta}{N}\sigma(2\bar{S}-1)\hat{C}_{SI}\end{aligned} \tag{A.3}$$

$$\begin{aligned}\frac{dC_{II}}{dt} &= \frac{\beta}{N}\bar{S}\bar{I} + \epsilon\bar{S} + (\gamma+\mu)\bar{I} + 2\left(\frac{\beta}{N}(1-\sigma)\bar{S} - (\gamma+\mu)\right)C_{II} \\ &\quad + \left(\frac{\beta}{N}(1-\sigma)(2\bar{I}+1) + 2\frac{\beta}{N}\sigma\bar{I} + 2\epsilon\right)C_{SI} + \frac{\beta}{N}\sigma S\hat{C}_{II} + \frac{\beta}{N}\sigma\hat{C}_{SI}\end{aligned} \tag{A.4}$$

$$\begin{aligned}\frac{dC_{SI}}{dt} &= -\frac{\beta}{N}\bar{S}\bar{I} - \epsilon\bar{S} - \mu\bar{I} + \left(\frac{\beta}{N}\bar{I} + \epsilon\right)C_{SS} - \frac{\beta}{N}(1-\sigma)\bar{S}C_{II} \\ &\quad + \left(\frac{\beta}{N}(1-\sigma)(\bar{S}-\bar{I}-1) - \frac{\beta}{N}\sigma\bar{I} - \epsilon - \gamma - 2\mu\right)C_{SI} - \frac{\beta}{N}\sigma\bar{S}\hat{C}_{II} \\ &\quad + \frac{\beta}{N}\sigma(\bar{S}-1)\hat{C}_{SI},\end{aligned} \tag{A.5}$$

and the ODEs for the three second-order moments are:

$$\frac{d\hat{C}_{SS}}{dt} = -2\frac{\beta}{N}\sigma\bar{S}C_{SI} - 2\left(\frac{\beta}{N}\bar{I} + \epsilon + \mu\right) - 2\frac{\beta}{N}(1-\sigma)\bar{S}\hat{C}_{SI} \tag{A.6}$$

$$\frac{d\hat{C}_{II}}{dt} = 2\frac{\beta}{N}\sigma\bar{S}C_{II} + 2\left(\frac{\beta}{N}(1-\sigma)\bar{S} - \gamma - \mu\right)\hat{C}_{II} + 2\left(\frac{\beta}{N}\bar{I} + \epsilon\right)\hat{C}_{SI} \tag{A.7}$$

$$\begin{aligned}\frac{d\hat{C}_{SI}}{dt} = &-\frac{\beta}{N}\sigma\bar{S}C_{II} + \frac{\beta}{N}\sigma\bar{S}C_{SI} + \left(\frac{\beta}{N}\bar{I} + \epsilon\right)\hat{C}_{SS} - \frac{\beta}{N}(1-\sigma)\bar{S}\hat{C}_{II} \\ &+ \left(\frac{\beta}{N}(1-\sigma)(\bar{S}-\bar{I}) - \epsilon - \gamma - 2\mu\right)\hat{C}_{SI}.\end{aligned} \tag{A.8}$$

To write down this set of ODEs, we make a second-order multivariate normal moment closure approximation and assume that third- and higher-order cumulants are equal to zero and thus third-order moments can be written in terms of the mean and (co)variance. For example, the third-order cumulant can be written as

$$\mathbb{E}[(S_1-\bar{S})(I_1-\bar{I})(I_2-\bar{I})] = \mathbb{E}[S_1I_1I_2] - \bar{S}\mathbb{E}[I_1I_2] - \bar{I}\mathbb{E}[S_1I_1] - \bar{I}\mathbb{E}[S_1I_2] + 2\bar{S}\bar{I}^2 \tag{A.9}$$

$$= \mathbb{E}[S_1I_1I_2] - \bar{S}\hat{C}_{II} - \bar{I}C_{SI} - \bar{I}\hat{C}_{SI} - \bar{S}\bar{I}^2, \tag{A.10}$$

and thus if we assume that $\mathbb{E}[(S_1-\bar{S})(I_1-\bar{I})(I_2-\bar{I})] = 0$ then the third-order moment can be written as

$$\mathbb{E}[S_1I_1I_2] = \bar{S}\hat{C}_{II} + \bar{I}C_{SI} + \bar{I}\hat{C}_{SI} + \bar{S}\bar{I}^2. \tag{A.11}$$

Analogous results hold for other third-order moments. We now derive the system of ODEs given by equations (A.1)-(A.8), making multivariate normal moment closure approximations where necessary. First we write down the ODEs for the first-order moments $\bar{S}$ and $\bar{I}$. For $\bar{S}$ we have

$$\frac{d\bar{S}}{dt} = \mathbb{E}\left[\left(\frac{\beta}{N}(1-\sigma)S_1I_1 + \frac{\beta}{N}\sigma S_1I_2 + \epsilon S_1\right)(-1) + \mu I_1(+1) + \mu(N-S_1-I_1)(+1)\right] \tag{A.12}$$

$$= \mathbb{E}\left[\mu N - \frac{\beta}{N}(1-\sigma)S_1I_1 - \frac{\beta}{N}\sigma S_1I_2 - \epsilon S_1 - \mu S_1\right] \tag{A.13}$$

$$= \mu N - \frac{\beta}{N}(1-\sigma)\mathbb{E}[S_1I_1] - \frac{\beta}{N}\sigma\mathbb{E}[S_1I_2] - \epsilon\mathbb{E}[S_1] - \mu\mathbb{E}[S_1] \tag{A.14}$$

$$= \mu N - \frac{\beta}{N}(1-\sigma)(C_{SI} + \bar{S}\bar{I}) - \frac{\beta}{N}\sigma(\hat{C}_{SI} + \bar{S}\bar{I}) - \epsilon\bar{S} - \mu\bar{S}. \tag{A.15}$$

Similarly, for $\bar{I}$ we have

$$\frac{d\bar{I}}{dt} = \mathbb{E}\left[\left(\frac{\beta}{N}(1-\sigma)S_1I_1 + \frac{\beta}{N}\sigma S_1I_2 + \epsilon S_1\right)(+1) + \gamma I_1(-1) + \mu I_1(-1)\right] \tag{A.16}$$

$$= \mathbb{E}\left[\frac{\beta}{N}(1-\sigma)S_1I_1 + \frac{\beta}{N}\sigma S_1I_2 + \epsilon S_1 - \gamma I_1 - \mu I_1\right] \tag{A.17}$$

$$= \frac{\beta}{N}(1-\sigma)\mathbb{E}[S_1I_1] + \frac{\beta}{N}\sigma\mathbb{E}[S_1I_2] + \epsilon\mathbb{E}[S_1] - \gamma\mathbb{E}[I_1] - \mu\mathbb{E}[I_1] \tag{A.18}$$

$$= \frac{\beta}{N}(1-\sigma)(C_{SI} + \bar{S}\bar{I}) + \frac{\beta}{N}\sigma(\hat{C}_{SI} + \bar{S}\bar{I}) + \epsilon\bar{S} - \gamma\bar{I} - \mu\bar{I}. \tag{A.19}$$

For C_{SS} we have

$$\frac{dC_{SS}}{dt} = \frac{d}{dt}\left(\mathbb{E}[S_1^2] - \bar{S}^2\right) = \frac{d\mathbb{E}[S_1^2]}{dt} - 2\bar{S}\frac{d\bar{S}}{dt}, \tag{A.20}$$

where

$$\frac{d\mathbb{E}[S_1^2]}{dt} = \mathbb{E}\Bigg[\left(\frac{\beta}{N}(1-\sigma)S_1I_1 + \frac{\beta}{N}\sigma S_1I_2 + \epsilon S_1\right)(-2S_1+1) + \mu I_1(2S_1+1) + \mu(N - S_1 - I_1)(2S_1+1)\Bigg] \tag{A.21}$$

$$= \mu N + \frac{\beta}{N}(1-\sigma)\mathbb{E}[S_1I_1] + \frac{\beta}{N}\sigma\mathbb{E}[S_1I_2] + \epsilon\mathbb{E}[S_1] - \mu\mathbb{E}[S_1] + 2\mu N\mathbb{E}[S_1] - 2\frac{\beta}{N}(1-\sigma)\mathbb{E}[S_1^2I_1] - 2\frac{\beta}{N}\sigma\mathbb{E}[S_1^2I_2] - 2\epsilon\mathbb{E}[S_1^2] - 2\mu\mathbb{E}[S_1^2] \tag{A.22}$$

and

$$2\bar{S}\frac{d\bar{S}}{dt} = 2\mu N\bar{S} - 2\frac{\beta}{N}(1-\sigma)(\bar{S}C_{SI} + \bar{S}^2\bar{I}) - 2\frac{\beta}{N}\sigma(\bar{S}\hat{C}_{SI} + \bar{S}^2\bar{I}) - 2\epsilon\bar{S}^2 - 2\mu\bar{S}^2. \tag{A.23}$$

For dC_{II}/dt we have

$$\frac{dC_{II}}{dt} = \frac{d\mathbb{E}[I_1^2]}{dt} - 2\bar{I}\frac{d\bar{I}}{dt}, \tag{A.24}$$

where

$$\frac{d\mathbb{E}[I_1^2]}{dt} = \mathbb{E}\left[\left(\frac{\beta}{N}(1-\sigma)S_1I_1 + \frac{\beta}{N}\sigma S_1I_2 + \epsilon S_1\right)(2I_1+1) + \gamma I_1(-2I_1+1) + \mu I_1(-2I_1+1)\right] \tag{A.25}$$

$$\begin{aligned} &= \frac{\beta}{N}(1-\sigma)\mathbb{E}[S_1I_1] + \frac{\beta}{N}\sigma\mathbb{E}[S_1I_2] + \epsilon\mathbb{E}[S_1] + \gamma\mathbb{E}[I_1] + \mu\mathbb{E}[I_1] \\ &\quad + 2\frac{\beta}{N}(1-\sigma)\mathbb{E}[S_1I_1^2] + 2\frac{\beta}{N}\sigma\mathbb{E}[S_1I_1I_2] + 2\epsilon\mathbb{E}[S_1I_1] - 2\gamma\mathbb{E}[I_1^2] - 2\mu\mathbb{E}[I_1I_1] \end{aligned} \tag{A.26}$$

and

$$2\bar{I}\frac{d\bar{I}}{dt} = 2\frac{\beta}{N}(1-\sigma)(\bar{I}C_{SI} + \bar{S}\bar{I}^2) + 2\frac{\beta}{N}\sigma(\bar{I}\hat{C}_{SI} + \bar{S}\bar{I}^2) + 2\epsilon\bar{S}\bar{I} - 2\gamma\bar{I}^2 - 2\mu\bar{I}^2. \tag{A.27}$$

For dC_{SI}/dt we have

$$\frac{dC_{SI}}{dt} = \frac{d\mathbb{E}[S_1I_2]}{dt} - \bar{I}\frac{d\bar{S}}{dt} - \bar{S}\frac{d\bar{I}}{dt}, \tag{A.28}$$

where

$$\begin{aligned} \frac{d\mathbb{E}[S_1I_1]}{dt} &= \mathbb{E}\left[\left(\frac{\beta}{N}(1-\sigma)S_1I_1 + \frac{\beta}{N}\sigma S_1I_2 + \epsilon S_1\right)(S_1 - I_1 - 1) + \gamma I_1(-S_1)\right. \\ &\quad \left. + \mu I_1(I_1 - S_1 - 1) + \mu(N - S_1 - I_1)(+I_1)\right] \end{aligned} \tag{A.29}$$

$$\begin{aligned} &= -\frac{\beta}{N}(1-\sigma)\mathbb{E}[S_1I_1] - \frac{\beta}{N}\sigma\mathbb{E}[S_1I_2] - \epsilon\mathbb{E}[S_1] - \mu\mathbb{E}[I_1] \\ &\quad + \mu N\mathbb{E}[I_1] - \frac{\beta}{N}(1-\sigma)\mathbb{E}[S_1I_1^2] - \frac{\beta}{N}\sigma\mathbb{E}[S_1I_1I_2] - \epsilon\mathbb{E}[S_1I_1] - \mu\mathbb{E}[S_1I_1] \\ &\quad + \frac{\beta}{N}(1-\sigma)\mathbb{E}[S_1^2I_1] + \frac{\beta}{N}\sigma\mathbb{E}[S_1^2I_2] + \epsilon\mathbb{E}[S_1^2] - \gamma\mathbb{E}[S_1I_1] - \mu\mathbb{E}[S_1I_1] \end{aligned} \tag{A.30}$$

and

$$\begin{aligned} \bar{I}\frac{d\bar{S}}{dt} + \bar{S}\frac{d\bar{I}}{dt} &= \mu N\bar{I} - \frac{\beta}{N}(1-\sigma)(\bar{I}C_{SI} + \bar{S}\bar{I}^2) - \frac{\beta}{N}\sigma(\bar{I}\hat{C}_{SI} + \bar{S}\bar{I}^2) - \epsilon\bar{S}\bar{I} - \mu\bar{S}\bar{I} \\ &\quad + \frac{\beta}{N}(1-\sigma)(\bar{S}C_{SI} + \bar{S}^2\bar{I}) + \frac{\beta}{N}\sigma(\bar{S}\hat{C}_{SI} + \bar{S}^2\bar{I}) + \epsilon\bar{S}^2 - \gamma\bar{S}\bar{I} - \mu\bar{S}\bar{I}. \end{aligned} \tag{A.31}$$

For $\hat{C}_{SS}$ we have

$$\hat{C}_{SS} = \frac{d\mathbb{E}[S_1 S_2]}{dt} - 2\bar{S}\frac{d\bar{S}}{dt}, \tag{A.32}$$

where

$$\begin{aligned}
\frac{d\mathbb{E}[S_1 S_2]}{dt} &= \mathbb{E}\Bigg[\left(\frac{\beta}{N}(1-\sigma)S_1 I_1 + \frac{\beta}{N}\sigma S_1 I_2 + \epsilon S_1\right)(-S_2) \\
&\quad + \mu I_1(+S_2) + \mu(N - S_1 - I_1)(+S_2) \\
&\quad + \left(\frac{\beta}{N}(1-\sigma)S_2 I_2 + \frac{\beta}{N}\sigma S_2 I_1 + \epsilon S_2\right)(-S_1) \\
&\quad + \mu I_2(+S_1) + \mu(N - S_2 - I_2)(+S_1)\Bigg] \qquad \text{(A.33)} \\
&= 2\mathbb{E}\Bigg[\left(\frac{\beta}{N}(1-\sigma)S_1 I_1 + \frac{\beta}{N}\sigma S_1 I_2 + \epsilon S_1\right)(-S_2) + \mu I_1(+S_2) \\
&\quad + \mu(N - S_1 - I_1)(+S_2)\Bigg] \qquad \text{(A.34)} \\
&= 2\mu N\bar{S} - 2\frac{\beta}{N}(1-\sigma)\mathbb{E}[S_1 S_2 I_1] - 2\frac{\beta}{N}\sigma\mathbb{E}[S_1 S_2 I_2] - 2\epsilon\mathbb{E}[S_1 S_2] - 2\mu\mathbb{E}[S_1 S_2], \qquad \text{(A.35)}
\end{aligned}$$

and

$$2\bar{S}\frac{d\bar{S}}{dt} = 2\mu N\bar{S} - 2\frac{\beta}{N}(1-\sigma)(\bar{S}C_{SI} + \bar{S}^2\bar{I}) - 2\frac{\beta}{N}\sigma(\bar{S}\hat{C}_{SI} + \bar{S}^2\bar{I}) - 2\epsilon\bar{S}^2 - 2\mu\bar{S}^2. \tag{A.36}$$

For $\hat{C}_{II}$ we have

$$\hat{C}_{II} = \frac{d\mathbb{E}[I_1 I_2]}{dt} - 2\bar{I}\frac{d\bar{I}}{dt}, \tag{A.37}$$

where

$$\frac{d\mathbb{E}[I_1I_2]}{dt} = \mathbb{E}\left[\left(\frac{\beta}{N}(1-\sigma)S_1I_1 + \frac{\beta}{N}\sigma S_1I_2 + \epsilon S_1\right)(+I_2) + \gamma I_1(-I_2) + \mu I_1(-I_2)\right.$$
$$\left. + \left(\frac{\beta}{N}(1-\sigma)S_2I_2 + \frac{\beta}{N}\sigma S_2I_1 + \epsilon S_2\right)(+I_1) + \gamma I_2(-I_1) + \mu I_2(-I_1)\right] \quad \text{(A.38)}$$
$$= 2\mathbb{E}\left[\left(\frac{\beta}{N}(1-\sigma)S_1I_1 + \frac{\beta}{N}\sigma S_1I_2 + \epsilon S_1\right)(+I_2) + \gamma I_1(-I_2) + \mu I_1(-I_2)\right] \quad \text{(A.39)}$$
$$= 2\frac{\beta}{N}(1-\sigma)\mathbb{E}[S_1I_1I_2] + 2\frac{\beta}{N}\sigma\mathbb{E}[S_1I_2^2] + 2\epsilon\mathbb{E}[S_1I_2] - 2\gamma\mathbb{E}[I_1I_2] - 2\mu\mathbb{E}[I_1I_2] \quad \text{(A.40)}$$

and

$$2\bar{I}\frac{d\bar{I}}{dt} = 2\frac{\beta}{N}(1-\sigma)(\bar{I}C_{SI} + \bar{S}\bar{I}^2) + 2\frac{\beta}{N}\sigma(\bar{I}\hat{C}_{SI} + \bar{S}\bar{I}^2) + 2\epsilon\bar{S}\bar{I} - 2\gamma\bar{I}^2 - 2\mu\bar{I}^2. \quad \text{(A.41)}$$

For $\hat{C}_{SI}$ we have

$$\hat{C}_{SI} = \frac{d\mathbb{E}[S_1I_2]}{dt} - \bar{I}\frac{d\bar{S}}{dt} - \bar{S}\frac{d\bar{I}}{dt}, \quad \text{(A.42)}$$

where

$$\frac{d\mathbb{E}[S_1I_2]}{dt} = \mathbb{E}\Bigg[\left(\frac{\beta}{N}(1-\sigma)S_1I_1 + \frac{\beta}{N}\sigma S_1I_2 + \epsilon S_1\right)(-I_2)$$
$$+ \mu I_1(+I_2) + \mu(N - S_1 - I_1)(+I_2)$$
$$+ \left(\frac{\beta}{N}(1-\sigma)S_2I_2 + \frac{\beta}{N}\sigma S_2I_1 + \epsilon S_2\right)(+S_1) + \gamma I_2(-S_1) + \mu I_2(-S_1)\Bigg] \quad \text{(A.43)}$$
$$= \mu N\mathbb{E}[I_2] - \frac{\beta}{N}(1-\sigma)\mathbb{E}[S_1I_1I_2] - \frac{\beta}{N}\sigma\mathbb{E}[S_1I_2^2] - \epsilon\mathbb{E}[S_1I_2] - \mu\mathbb{E}[S_1I_2]$$
$$+ \frac{\beta}{N}(1-\sigma)\mathbb{E}[S_1S_2I_2] + \frac{\beta}{N}\sigma\mathbb{E}[S_1S_2I_1] + \epsilon\mathbb{E}[S_1S_2] - \gamma\mathbb{E}[S_1I_2] - \mu\mathbb{E}[S_1I_2] \quad \text{(A.44)}$$

and

$$\bar{I}\frac{d\bar{S}}{dt} + \bar{S}\frac{d\bar{I}}{dt} = \mu N\bar{I} - \frac{\beta}{N}(1-\sigma)(\bar{I}C_{SI} + \bar{S}\bar{I}^2) - \frac{\beta}{N}\sigma(\bar{I}\hat{C}_{SI} + \bar{S}\bar{I}^2) - \epsilon\bar{S}\bar{I} - \mu\bar{S}\bar{I}$$
$$+ \frac{\beta}{N}(1-\sigma)(\bar{S}C_{SI} + \bar{S}^2\bar{I}) + \frac{\beta}{N}\sigma(\bar{S}\hat{C}_{SI} + \bar{S}^2\bar{I}) + \epsilon\bar{S}^2 - \gamma\bar{S}\bar{I} - \mu\bar{S}\bar{I}. \tag{A.45}$$

Appendix B

Appendix to Chapter 4

B.1 The ODE system approximating the stochastic endemic infection model on the complete network

For the stochastic epidemic metapopulation model on the complete network with P populations, where the coupling between interacting populations is $\sigma \in [0, 1/k], k = P - 1$, we can approximate the stochastic process by the following system of 8 ODEs. There are 5 equations for the within-population moments, of which two are first-order:

$$\frac{d\bar{S}}{dt} = \mu N - \frac{\beta}{N}(1 - k\sigma)(C_{SI} + \bar{S}\bar{I}) - k\frac{\beta}{N}\sigma(\hat{C}_{SI} + \bar{S}\bar{I}) - \epsilon\bar{S} - \mu\bar{S} \tag{B.1}$$

$$\frac{d\bar{I}}{dt} = \frac{\beta}{N}(1 - k\sigma)(C_{SI} + \bar{S}\bar{I}) + k\frac{\beta}{N}\sigma(\hat{C}_{SI} + \bar{S}\bar{I}) + \epsilon\bar{S} - \gamma\bar{I} - \mu\bar{I}, \tag{B.2}$$

and three are second-order:

$$\frac{dC_{SS}}{dt} = \mu N\frac{\beta}{N}\bar{S}\bar{I} + \epsilon\bar{S} - \mu\bar{S} - 2\left(\frac{\beta}{N}\bar{I} + \epsilon + \mu\right) - \frac{\beta}{N}(1-k\sigma)(2\bar{S}-1) + k\frac{\beta}{N}\sigma(2\bar{S}-1)\hat{C}_{SI} \tag{B.3}$$

$$\frac{dC_{II}}{dt} = \frac{\beta}{N}\bar{S}\bar{I} + \epsilon\bar{S} + (\gamma+\mu)\bar{I} + 2\left(\frac{\beta}{N}(1-k\sigma)\bar{S} - (\gamma+\mu)\right)C_{II} + \left(\frac{\beta}{N}(1-k\sigma)(2\bar{I}+1) + 2k\frac{\beta}{N}\sigma\bar{I} + 2\epsilon\right)C_{SI} + 2k\frac{\beta}{N}\sigma\bar{S}\hat{C}_{II} + k\frac{\beta}{N}\sigma\hat{C}_{SI} \tag{B.4}$$

$$\frac{dC_{SI}}{dt} = -\frac{\beta}{N}\bar{S}\bar{I} - \epsilon\bar{S} - \mu\bar{I} + \left(\frac{\beta}{N}\bar{I} + \epsilon\right)C_{SS} - \frac{\beta}{N}(1-k\sigma)\bar{S}C_{II} + \left(\frac{\beta}{N}(1-k\sigma)(\bar{S}-\bar{I}-1) - k\frac{\beta}{N}\sigma\bar{I} - \epsilon - \gamma - 2\mu\right)C_{SI} - k\frac{\beta}{N}\sigma\bar{S}\hat{C}_{II} + k\frac{\beta}{N}\sigma(\bar{S}-1)\hat{C}_{SI}. \tag{B.5}$$

In addition, there are 3 equations for the between-population moments:

$$\frac{d\hat{C}_{SS}}{dt} = -2\frac{\beta}{N}\sigma\bar{S}C_{SI} - 2\left(\frac{\beta}{N}\bar{I} + \epsilon + \mu\right)\hat{C}_{SS} - 2\left(\frac{\beta}{N}(1-k\sigma)\bar{S} + (k-1)\frac{\beta}{N}\sigma\bar{S}\right)\hat{C}_{SI} \tag{B.6}$$

$$\frac{d\hat{C}_{II}}{dt} = 2\frac{\beta}{N}\sigma\bar{S}C_{II} + 2\left(\frac{\beta}{N}(1-k\sigma)\bar{S} + (k-1)\frac{\beta}{N}\sigma\bar{S} - \gamma - \mu\right)\hat{C}_{II} + 2\left(\frac{\beta}{N}\bar{I} + \epsilon\right)\hat{C}_{SI} \tag{B.7}$$

$$\frac{d\hat{C}_{SI}}{dt} = -\frac{\beta}{N}\sigma\bar{S}C_{II} + \frac{\beta}{N}\sigma\bar{S}C_{SI} + \left(\frac{\beta}{N}\bar{I} + \epsilon\right)\hat{C}_{SS} - \left(\frac{\beta}{N}(1-k\sigma) + (k-1)\frac{\beta}{N}\sigma\bar{S}\right)\hat{C}_{II} + \left(\frac{\beta}{N}(1-k\sigma)(\bar{S}-\bar{I}) + \frac{\beta}{N}\sigma((k-1)\bar{S} - k\bar{I}) - \epsilon - \gamma - 2\mu\right)\hat{C}_{SI}. \tag{B.8}$$

B.2 Derivation of the approximation for the complete network

For the complete network on P populations, where the coupling between interacting populations is $\sigma \in [0, 1/k], k = P - 1$, we can show that the correlation, ρ, between the number of infected individuals in any pair of populations is equal to

$$\rho = \frac{\sigma}{\xi + \sigma} - \Delta, \tag{B.9}$$

where

$$\xi = \frac{N(\gamma + \mu) - \beta \bar{S}^*}{\beta \bar{S}^*} \tag{B.10}$$

and

$$\Delta = \frac{(\beta \bar{I}^* + N\epsilon)\frac{\hat{C}^*_{SI}}{C^*_{II}}}{\beta(1-\sigma)\bar{S}^* - N(\gamma + \mu)}. \tag{B.11}$$

To derive this result we begin with the moment equation for $\hat{C}_{II}$:

$$\frac{d\hat{C}_{II}}{dt} = 2\frac{\beta}{N}\sigma \bar{S} C_{II} + 2\left(\frac{\beta}{N}(1-\sigma)\bar{S} - \gamma - \mu\right)\hat{C}_{II} + 2\left(\frac{\beta}{N}\bar{I} + \epsilon\right)\hat{C}_{SI}. \tag{B.12}$$

At equilibrium, $d\hat{C}_{II}/dt = 0$ and if we divide by $2C^*_{II}/N$, then

$$0 = \beta\sigma\bar{S}^* + \left(\beta(1-\sigma)\bar{S}^* - N(\gamma+\mu)\right)\rho + (\beta\bar{I}^* + N\epsilon)\frac{\hat{C}^*_{SI}}{C^*_{II}}, \tag{B.13}$$

and hence we have the following approximation for the correlation:

$$\begin{aligned}
\rho &= \frac{-\beta\sigma\bar{S}^*}{\beta(1-\sigma)\bar{S}^* - N(\gamma+\mu)} - \frac{\beta\bar{I}^* + N\epsilon}{\beta(1-\sigma)\bar{S}^* - N(\gamma+\mu)}\frac{\hat{C}^*_{SI}}{C^*_{II}} \\
&= \frac{\sigma}{\frac{N(\gamma+\mu)-\beta\bar{S}^*}{\beta\bar{S}^*} + \sigma} - \Delta \\
&= \frac{\sigma}{\xi + \sigma} - \Delta.
\end{aligned} \tag{B.14}$$

B.3 The ODE system approximating the stochastic endemic infection model on the tree network

B.3.1 The full k-regular tree network

We can approximate the stochastic epidemic process on the full k-regular tree network, where the coupling between interacting populations is $\sigma \in [0, 1/k]$, by the following system of ODES. There are 5 equations for the within-population moments, of which two are first-order:

$$\frac{d\bar{S}}{dt} = \mu N - \frac{\beta}{N}(1-k\sigma)(C_{SI} + \bar{S}\bar{I}) - k\frac{\beta}{N}\sigma(\hat{C}^{(1)}_{SI} + \bar{S}\bar{I}) - \epsilon\bar{S} - \mu\bar{S} \tag{B.15}$$

$$\frac{d\bar{I}}{dt} = \frac{\beta}{N}(1-k\sigma)(C_{SI} + \bar{S}\bar{I}) + k\frac{\beta}{N}\sigma(\hat{C}^{(1)}_{SI} + \bar{S}\bar{I}) + \epsilon\bar{S} - \gamma\bar{I} - \mu\bar{I} \tag{B.16}$$

and three are second-order

$$\frac{dC_{SS}}{dt} = \mu N\frac{\beta}{N}\bar{S}\bar{I} + \epsilon\bar{S} - \mu\bar{S} - 2\left(\frac{\beta}{N}\bar{I} + \epsilon + \mu\right) - \frac{\beta}{N}(1-k\sigma)(2\bar{S}-1) + k\frac{\beta}{N}\sigma(2\bar{S}-1)\hat{C}^{(1)}_{SI} \tag{B.17}$$

$$\begin{aligned}\frac{dC_{II}}{dt} = {} & \frac{\beta}{N}\bar{S}\bar{I} + \epsilon\bar{S} + (\gamma+\mu)\bar{I} + 2\left(\frac{\beta}{N}(1-k\sigma)\bar{S} - (\gamma+\mu)\right)C_{II} \\ & + \left(\frac{\beta}{N}(1-k\sigma)(2\bar{I}+1) + 2k\frac{\beta}{N}\sigma\bar{I} + 2\epsilon\right)C_{SI} + 2k\frac{\beta}{N}\sigma\bar{S}\hat{C}^{(1)}_{II} + k\frac{\beta}{N}\sigma\hat{C}^{(1)}_{SI}\end{aligned} \tag{B.18}$$

$$\begin{aligned}\frac{dC_{SI}}{dt} = {} & -\frac{\beta}{N}\bar{S}\bar{I} - \epsilon\bar{S} - \mu\bar{I} + \left(\frac{\beta}{N}\bar{I} + \epsilon\right)C_{SS} - \frac{\beta}{N}(1-k\sigma)\bar{S}C_{II} \\ & + \left(\frac{\beta}{N}(1-k\sigma)(\bar{S}-\bar{I}-1) - k\frac{\beta}{N}\sigma\bar{I} - \epsilon - \gamma - 2\mu\right)C_{SI} - k\frac{\beta}{N}\sigma\bar{S}\hat{C}^{(1)}_{II} + k\frac{\beta}{N}\sigma(\bar{S}-1)\hat{C}^{(1)}_{SI}.\end{aligned} \tag{B.19}$$

In addition, there are $3d$ equations for the between-population moments, $d \geq 1$:

$$\frac{d\hat{C}_{SS}^{(d)}}{dt} = -2\frac{\beta}{N}\sigma\bar{S}\hat{C}_{SI}^{(d-1)} - 2\left(\frac{\beta}{N}\bar{I} + \epsilon + \mu\right)\hat{C}_{SS}^{(d)} - 2\frac{\beta}{N}(1-k\sigma)\bar{S}\hat{C}_{SI}^{(d)} - 2(k-1)\frac{\beta}{N}\sigma\bar{S}\hat{C}_{SI}^{(d+1)} \tag{B.20}$$

$$\frac{d\hat{C}_{II}^{(d)}}{dt} = 2\frac{\beta}{N}\sigma\bar{S}\hat{C}_{II}^{(d-1)} + 2\left(\frac{\beta}{N}(1-k\sigma)\bar{S} - \gamma - \mu\right)\hat{C}_{II}^{(d)} + 2\left(\frac{\beta}{N}\bar{I} + \epsilon\right)\hat{C}_{SI}^{(d)} + 2(k-1)\frac{\beta}{N}\sigma\bar{S}\hat{C}_{II}^{(d+1)} \tag{B.21}$$

$$\begin{aligned}\frac{d\hat{C}_{SI}^{(d)}}{dt} = &-\frac{\beta}{N}\sigma\bar{S}C_{II}^{(d-1)} + \frac{\beta}{N}\sigma\bar{S}C_{SI}^{(d-1)} + \left(\frac{\beta}{N}\bar{I} + \epsilon\right)\hat{C}_{SS}^{(d)} - \frac{\beta}{N}(1-k\sigma)\hat{C}_{II}^{(d)} \\ &+ \left(\frac{\beta}{N}(1-k\sigma)(\bar{S}-\bar{I}) - k\frac{\beta}{N}\sigma\bar{I} - \epsilon - \gamma - 2\mu\right)\hat{C}_{SI}^{(d)} - (k-1)\frac{\beta}{N}\sigma\bar{S}\hat{C}_{II}^{(d+1)} + (k-1)\frac{\beta}{N}\sigma\bar{S}\hat{C}_{SI}^{(d+1)},\end{aligned} \tag{B.22}$$

and where $\hat{C}_{XY}^{(0)} = C_{XY}$.

B.3.2 The D-truncated k-regular tree network

We can approximate the stochastic epidemic process on the D-truncated k-regular tree network, where the coupling between interacting populations is $\sigma \in [0, 1/k]$, by the following system of $3D + 5$ ODEs. There are 5 equations for the within-population moments, of which two are first-order:

$$\frac{d\bar{S}}{dt} = \mu N - \frac{\beta}{N}(1-k\sigma)(C_{SI} + \bar{S}\bar{I}) - k\frac{\beta}{N}\sigma(\hat{C}_{SI}^{(1)} + \bar{S}\bar{I}) - \epsilon\bar{S} - \mu\bar{S} \tag{B.23}$$

$$\frac{d\bar{I}}{dt} = \frac{\beta}{N}(1-k\sigma)(C_{SI} + \bar{S}\bar{I}) + k\frac{\beta}{N}\sigma(\hat{C}_{SI}^{(1)} + \bar{S}\bar{I}) + \epsilon\bar{S} - \gamma\bar{I} - \mu\bar{I} \tag{B.24}$$

and three are second-order:

$$\frac{dC_{SS}}{dt} = \mu N\frac{\beta}{N}\bar{S}\bar{I} + \epsilon\bar{S} - \mu\bar{S} - 2\left(\frac{\beta}{N}\bar{I} + \epsilon + \mu\right) - \frac{\beta}{N}(1-k\sigma)(2\bar{S}-1) + k\frac{\beta}{N}\sigma(2\bar{S}-1)\hat{C}_{SI}^{(1)} \tag{B.25}$$

$$\begin{aligned}\frac{dC_{II}}{dt} &= \frac{\beta}{N}\bar{S}\bar{I} + \epsilon\bar{S} + (\gamma+\mu)\bar{I} + 2\left(\frac{\beta}{N}(1-k\sigma)\bar{S} - (\gamma+\mu)\right)C_{II} \\ &\quad + \left(\frac{\beta}{N}(1-k\sigma)(2\bar{I}+1) + 2k\frac{\beta}{N}\sigma\bar{I} + 2\epsilon\right)C_{SI} + 2k\frac{\beta}{N}\sigma\bar{S}\hat{C}_{II}^{(1)} + k\frac{\beta}{N}\sigma\hat{C}_{SI}^{(1)}\end{aligned} \tag{B.26}$$

$$\begin{aligned}\frac{dC_{SI}}{dt} &= -\frac{\beta}{N}\bar{S}\bar{I} - \epsilon\bar{S} - \mu\bar{I} + \left(\frac{\beta}{N}\bar{I} + \epsilon\right)C_{SS} - \frac{\beta}{N}(1-k\sigma)\bar{S}C_{II} \\ &\quad + \left(\frac{\beta}{N}(1-k\sigma)(\bar{S}-\bar{I}-1) - k\frac{\beta}{N}\sigma\bar{I} - \epsilon - \gamma - 2\mu\right)C_{SI} - k\frac{\beta}{N}\sigma\bar{S}\hat{C}_{II}^{(1)} + k\frac{\beta}{N}\sigma(\bar{S}-1)\hat{C}_{SI}^{(1)}.\end{aligned} \tag{B.27}$$

In addition, there are $3D$ equations for the between-population moments; for $d = 1, \ldots, D-1$ we have:

$$\frac{d\hat{C}_{SS}^{(d)}}{dt} = -2\frac{\beta}{N}\sigma\bar{S}\hat{C}_{SI}^{(d-1)} - 2\left(\frac{\beta}{N}\bar{I} + \epsilon + \mu\right)\hat{C}_{SS}^{(d)} - 2\frac{\beta}{N}(1-k\sigma)\bar{S}\hat{C}_{SI}^{(d)} - 2(k-1)\frac{\beta}{N}\sigma\bar{S}\hat{C}_{SI}^{(d+1)} \tag{B.28}$$

$$\frac{d\hat{C}_{II}^{(d)}}{dt} = 2\frac{\beta}{N}\sigma\bar{S}\hat{C}_{II}^{(d-1)} + 2\left(\frac{\beta}{N}(1-k\sigma)\bar{S} - \gamma - \mu\right)\hat{C}_{II}^{(d)} + 2\left(\frac{\beta}{N}\bar{I} + \epsilon\right)\hat{C}_{SI}^{(d)} + 2(k-1)\frac{\beta}{N}\sigma\bar{S}\hat{C}_{II}^{(d+1)} \tag{B.29}$$

$$\begin{aligned}\frac{d\hat{C}_{SI}^{(d)}}{dt} &= -\frac{\beta}{N}\sigma\bar{S}C_{II}^{(d-1)} + \frac{\beta}{N}\sigma\bar{S}C_{SI}^{(d-1)} + \left(\frac{\beta}{N}\bar{I} + \epsilon\right)\hat{C}_{SS}^{(d)} - \frac{\beta}{N}(1-k\sigma)\hat{C}_{II}^{(d)} \\ &\quad + \left(\frac{\beta}{N}(1-k\sigma)(\bar{S}-\bar{I}) - k\frac{\beta}{N}\sigma\bar{I} - \epsilon - \gamma - 2\mu\right)\hat{C}_{SI}^{(d)} - (k-1)\frac{\beta}{N}\sigma\bar{S}\hat{C}_{II}^{(d+1)} + (k-1)\frac{\beta}{N}\sigma\bar{S}\hat{C}_{SI}^{(d+1)},\end{aligned} \tag{B.30}$$

and for $d = D$ we have:

$$\frac{d\hat{C}_{SS}^{(D)}}{dt} = -2\frac{\beta}{N}\sigma\bar{S}\hat{C}_{SI}^{(D-1)} - 2\left(\frac{\beta}{N}\bar{I} + \epsilon + \mu\right)\hat{C}_{SS}^{(D)} - 2\frac{\beta}{N}(1-k\sigma)\bar{S}\hat{C}_{SI}^{(D)} \tag{B.31}$$

$$\frac{d\hat{C}_{II}^{(D)}}{dt} = 2\frac{\beta}{N}\sigma\bar{S}\hat{C}_{II}^{(D-1)} + 2\left(\frac{\beta}{N}(1-k\sigma)\bar{S} - \gamma - \mu\right)\hat{C}_{II}^{(D)} + 2\left(\frac{\beta}{N}\bar{I} + \epsilon\right)\hat{C}_{SI}^{(D)} \tag{B.32}$$

$$\begin{aligned}\frac{d\hat{C}_{SI}^{(D)}}{dt} = &-\frac{\beta}{N}\sigma\bar{S}C_{II}^{(D-1)} + \frac{\beta}{N}\sigma\bar{S}C_{SI}^{(D-1)} + \left(\frac{\beta}{N}\bar{I} + \epsilon\right)\hat{C}_{SS}^{(D)} - \frac{\beta}{N}(1-k\sigma)\hat{C}_{II}^{(D)} \\ &+ \left(\frac{\beta}{N}(1-k\sigma)(\bar{S}-\bar{I}) - k\frac{\beta}{N}\sigma\bar{I} - \epsilon - \gamma - 2\mu\right)\hat{C}_{SI}^{(D)}.\end{aligned} \tag{B.33}$$

B.4 Derivation of the approximation for the k-regular tree network

For the k-regular tree network, where the coupling between interacting populations is $\sigma \in [0, 1/k], k = P - 1$, we can show that the correlation, ρ, between the number of infected individuals in any pair of populations distance d apart is the solution to

$$\rho_d = \frac{\sigma}{\xi + k\sigma} \left(\rho_{d-1} + (k-1)\rho_{d+1}\right) - \Delta^{(d)}, \tag{B.34}$$

where

$$\xi = \frac{N(\gamma + \mu) - \beta \bar{S}^*}{\beta \bar{S}^*} \tag{B.35}$$

and

$$\Delta^{(d)} = \frac{(\beta \bar{I}^* + N\epsilon)}{\beta(1 - k\sigma)\bar{S}^* - N(\gamma + \mu)} \frac{\hat{C}_{SI}^{(d)*}}{C_{II}^*}. \tag{B.36}$$

We derive this result we begin with the moment equation $\hat{C}_{II}^{(d)}$:

$$\begin{aligned} \frac{\hat{C}_{II}^{(d)}}{dt} = {} & 2\frac{\beta}{N}\sigma \bar{S} \hat{C}_{II}^{(d-1)} + 2\left(\frac{\beta}{N}(1 - k\sigma)\bar{S} - (\gamma + \mu)\right)\hat{C}_{II}^{(d)} + 2\left(\frac{\beta}{N}\bar{I} + \epsilon\right)\hat{C}_{SI}^{(d)} \\ & + 2(k-1)\frac{\beta}{N}\sigma \bar{S} \hat{C}_{II}^{(d+1)} \end{aligned} \tag{B.37}$$

At equilibrium $d\hat{C}_{II}^{(d)}/dt = 0$ and if we divide by $2C_{II}^*/N$ then

$$\begin{aligned} 0 = {} & \beta\sigma \bar{S}^* \rho_{d-1} + (\beta(1 - k\sigma)\bar{S}^* - N(\gamma + \mu))\rho_d + \beta\sigma \bar{S}^*(k-1)\rho_{d+1} \\ & + (\beta \bar{I}^* + N\epsilon)\frac{\hat{C}_{SI}^{(d)*}}{C_{II}^*}, \end{aligned} \tag{B.38}$$

and hence we have the following relationship between ρ_{d-1}, ρ_d and ρ_{d+1}

$$\begin{aligned}\rho_d &= \frac{\beta\sigma\bar{S}^*}{N(\gamma+\mu)-\beta(1-k\sigma)\bar{S}^*}\left(\rho_{d-1}+(k-1)\rho_{d+1}\right) - \frac{\beta\bar{I}^*+N\epsilon}{N(\gamma+\mu)-\beta(1-k\sigma)\bar{S}^*}\frac{\hat{C}_{SI}^{(d)*}}{C_{II}^*} \\ &= \frac{\sigma}{\frac{N(\gamma+\mu)-\beta\bar{S}^*}{\beta\bar{S}^*}+k\sigma}\left(\rho_{d-1}+(k-1)\rho_{d+1}\right) - \Delta_k^{(d)} \\ &= \frac{\sigma}{\xi+k\sigma}\left(\rho_{d-1}+(k-1)\rho_{d+1}\right) - \Delta^{(d)}. \end{aligned} \tag{B.39}$$

B.5 The ODE system approximating the stochastic endemic infection model on the star network

For the stochastic epidemic process on the star network with k leaf populations, where the coupling between interacting populations is $\sigma \in [0, 1/k]$, by the following system of seventeen ODES. This system comprises ten equations for the within-population moments, of which five are for within the hub population:

$$\frac{d\bar{S}_H}{dt} = \mu N - \frac{\beta}{N}(1-k\sigma)(C^H_{SI} + \bar{S}_H\bar{I}_H) - k\frac{\beta}{N}\sigma(\hat{C}_{S_HI_L} + \bar{S}_h\bar{I}_L) - \epsilon\bar{S}_H - \mu\bar{S}_H \tag{B.40}$$

$$\frac{d\bar{I}_H}{dt} = \frac{\beta}{N}(1-k\sigma)(C^H_{SI} + \bar{S}_H\bar{I}_H) + k\frac{\beta}{N}\sigma(\hat{C}_{S_HI_L} + \bar{S}_H\bar{I}_L) + \epsilon\bar{S}_H - \gamma\bar{I}_H - \mu\bar{I}_H \tag{B.41}$$

$$\begin{aligned}\frac{dC^H_{SS}}{dt} &= \mu N + \frac{\beta}{N}(1-\sigma)\,\bar{S}_H\bar{I}_L + \epsilon\bar{S}_H - \mu\bar{S}_H - 2\left(\frac{\beta}{N}(1-k\sigma)\bar{I}_H + k\frac{\beta}{N}\sigma\bar{I}_L + \epsilon + \mu\right)C^H_{SS} \\ &\quad - \frac{\beta}{N}(1-k\sigma)(2\bar{S}_H - 1)C^H_{SI} - k\frac{\beta}{N}\sigma(2\bar{S}_H - 1)\hat{C}_{S_HI_L}\end{aligned} \tag{B.42}$$

$$\begin{aligned}\frac{dC^H_{II}}{dt} &= \frac{\beta}{N}(1-k\sigma)\bar{S}_H\bar{I}_H + k\frac{\beta}{N}\sigma\bar{S}_H\bar{I}_L + \epsilon\bar{S}_H + (\gamma+\mu)\bar{I}_H + 2\left(\frac{\beta}{N}(1-k\sigma)\bar{S}_H - (\gamma+\mu)\right)C^H_{II} \\ &\quad + \left(\frac{\beta}{N}(1-k\sigma)(2\bar{I}_H + 1) + 2k\frac{\beta}{N}\sigma\bar{I}_L + 2\epsilon\right)C^H_{SI} + 2k\frac{\beta}{N}\sigma\bar{S}_H\hat{C}^H_{II} + k\frac{\beta}{N}\sigma\hat{C}_{S_HI_L}\end{aligned} \tag{B.43}$$

$$\begin{aligned}\frac{dC^H_{SI}}{dt} &= -\frac{\beta}{N}(1-k\sigma)\bar{S}_H\bar{I}_H - k\frac{\beta}{N}\sigma\bar{S}_H\bar{I}_L - \epsilon\bar{S}_H - \mu\bar{I}_H + \left(\frac{\beta}{N}(1-k\sigma)\bar{I}_H + k\frac{\beta}{N}\sigma\bar{I}_L + \epsilon\right)C^H_{SS} \\ &\quad - \frac{\beta}{N}(1-k\sigma)\bar{S}_HC^H_{II} + \left(\frac{\beta}{N}(1-k\sigma)(\bar{S}_H - \bar{I}_H - 1) - k\frac{\beta}{N}\sigma\bar{I}_L - \epsilon - \gamma - 2\mu\right)C^H_{SI} \\ &\quad - k\frac{\beta}{N}\sigma\bar{S}_H\hat{C}^H_{II} + k\frac{\beta}{N}\sigma(\bar{S}_H - 1)\hat{C}_{S_HI_L}\end{aligned} \tag{B.44}$$

and the remaining five for within the leaf population

$$\frac{d\bar{S}_L}{dt} = \mu N - \frac{\beta}{N}(1-\sigma)(C^L_{SI} + \bar{S}_L\bar{I}_L) - \frac{\beta}{N}\sigma(\hat{C}_{S_LI_H} + \bar{S}_L\bar{I}_H) - \epsilon\bar{S}_L - \mu\bar{S}_L \tag{B.45}$$

$$\frac{d\bar{I}_L}{dt} = \frac{\beta}{N}(1-\sigma)(C^L_{SI} + \bar{S}_L\bar{I}_L) + \frac{\beta}{N}\sigma(\hat{C}_{S_LI_H} + \bar{S}_L\bar{I}_H) + \epsilon\bar{S}_L - \gamma\bar{I}_L - \mu\bar{I}_L \tag{B.46}$$

$$\begin{aligned}\frac{dC^L_{SS}}{dt} &= \mu N + \frac{\beta}{N}(1-\sigma)\bar{S}_L\bar{I}_L + \frac{\beta}{N}\sigma\bar{S}_L\bar{I}_H + \epsilon\bar{S}_H - \mu\bar{S}_H \\ &\quad - 2\left(\frac{\beta}{N}(1-\sigma)\bar{I}_L + \frac{\beta}{N}\sigma\bar{I}_H + \epsilon + \mu\right)C^L_{SS} \\ &\quad - \frac{\beta}{N}(1-\sigma)(2\bar{S}_L - 1)C^L_{SI} - \frac{\beta}{N}\sigma(2\bar{S}_L - 1)\hat{C}_{S_LI_H}\end{aligned} \tag{B.47}$$

$$\begin{aligned}\frac{dC^L_{II}}{dt} &= \frac{\beta}{N}(1-\sigma)\bar{S}_L\bar{I}_L + \frac{\beta}{N}\sigma\bar{S}_L\bar{I}_H + \epsilon\bar{S}_L + (\gamma+\mu)\bar{I}_L + 2\left(\frac{\beta}{N}(1-\sigma)\bar{S}_L - (\gamma+\mu)\right)C^L_{II} \\ &\quad + \left(\frac{\beta}{N}(1-\sigma)(2\bar{I}_L + 1) + 2\frac{\beta}{N}\sigma\bar{I}_H + 2\epsilon\right)C^L_{SI} + 2\frac{\beta}{N}\sigma\bar{S}_L\hat{C}^H_{II} + \frac{\beta}{N}\sigma\hat{C}_{S_LI_H}\end{aligned} \tag{B.48}$$

$$\begin{aligned}\frac{dC^L_{SI}}{dt} &= -\frac{\beta}{N}(1-\sigma)\bar{S}_L\bar{I}_L - \frac{\beta}{N}\sigma\bar{S}_L\bar{I}_H - \epsilon\bar{S}_L - \mu\bar{I}_L + \left(\frac{\beta}{N}(1-\sigma)\bar{I}_L + \frac{\beta}{N}\sigma\bar{I}_H + \epsilon\right)C^L_{SS} \\ &\quad - \frac{\beta}{N}(1-\sigma)\bar{S}_L C^L_{II} + \left(\frac{\beta}{N}(1-\sigma)(\bar{S}_L - \bar{I}_L - 1) + \frac{\beta}{N}\sigma\bar{I}_H - \epsilon - \gamma - 2\mu\right)C^L_{SI} \\ &\quad - \frac{\beta}{N}\sigma\bar{S}_L\hat{C}^H_{II} + k\frac{\beta}{N}\sigma(\bar{S}_L - 1)\hat{C}_{S_LH_H}.\end{aligned} \tag{B.49}$$

The remaining 7 equations are for the between-population moments:

$$\begin{aligned}\frac{d\hat{C}^H_{SS}}{dt} &= \frac{\beta}{N}\sigma\bar{S}_L C^H_{SI} - \frac{\beta}{N}\sigma\bar{S}_H C^L_{SI} \\ &\quad - \left(\frac{\beta}{N}(1-k\sigma)\bar{I}_H + k\frac{\beta}{N}\sigma\bar{I}_L + \frac{\beta}{N}(1-\sigma)\bar{I}_L + \frac{\beta}{N}\sigma\bar{I}_H + 2\epsilon + 2\mu\right)\hat{C}^H_{SS} \\ &\quad - \frac{\beta}{N}(1-\sigma)\bar{S}_L\hat{C}_{S_HI_L} - \frac{\beta}{N}(1-k\sigma)\bar{S}_H\hat{C}_{S_LI_H} - (k-1)\frac{\beta}{N}\sigma\bar{S}_H\hat{C}^L_{SI} \qquad \text{(B.50)}\end{aligned}$$

$$\frac{d\hat{C}^L_{SS}}{dt} = -2\left[\left(\frac{\beta}{N}(1-\sigma)\bar{I}_L + \frac{\beta}{N}\sigma\bar{I}_H + \epsilon + \mu\right)\hat{C}^L_{SS} + \frac{\beta}{N}\sigma\bar{S}_L\hat{C}_{S_LI_H} + \frac{\beta}{N}(1-\sigma)\bar{S}_L\hat{C}^L_{SI}\right] \qquad \text{(B.51)}$$

$$\begin{aligned}\frac{d\hat{C}^H_{II}}{dt} &= \frac{\beta}{N}\sigma\bar{S}_L C^H_{II} + \frac{\beta}{N}\sigma\bar{S}_H C^L_{II} + \left(\frac{\beta}{N}(1-k\sigma)\bar{S}_H + \frac{\beta}{N}(1-\sigma)\bar{S}_L - 2(\gamma+\mu)\right)\hat{C}^H_{II} \\ &\quad + (k-1)\frac{\beta}{N}\sigma\bar{S}_H\hat{C}^L_{II} \\ &\quad + \left(\frac{\beta}{N}(1-k\sigma)\bar{I}_H + k\frac{\beta}{N}\sigma\bar{I}_L + \epsilon\right)\hat{C}_{S_HI_L} + \left(\frac{\beta}{N}(1-\sigma)\bar{I}_L + \frac{\beta}{N}\sigma\bar{I}_H + \epsilon\right)\hat{C}_{S_LI_H} \qquad \text{(B.52)}\end{aligned}$$

$$\frac{d\hat{C}^L_{II}}{dt} = 2\left[\frac{\beta}{N}\sigma\bar{S}_L\hat{C}^H_{II} + \left(\frac{\beta}{N}(1-\sigma) - (\gamma+\mu)\right)\hat{C}^L_{II} + \left(\frac{\beta}{N}(1-\sigma)\bar{I}_L + \frac{\beta}{N}\sigma\bar{I}_H\epsilon\right)\hat{C}^L_{SI}\right] \qquad \text{(B.53)}$$

$$\begin{aligned}\frac{d\hat{C}_{S_HI_L}}{dt} &= -\frac{\beta}{N}\sigma\bar{S}_H C^L_{II} + \frac{\beta}{N}\sigma\bar{S}_L\bar{SI}^H + \left(\frac{\beta}{N}(1-\sigma)\bar{I}_L + \frac{\beta}{N}\sigma\bar{I}_H + \epsilon\right)\hat{C}^H_{SS} - \frac{\beta}{N}(1-k\sigma)\bar{S}_H\hat{C}^H_{II} \\ &\quad - (k-1)\frac{\beta}{N}\sigma\bar{S}_H\hat{C}^L_{II} + \left(\frac{\beta}{N}(1-\sigma)\bar{S}_L - \frac{\beta}{N}(1-k\sigma)\bar{I}_H - k\frac{\beta}{N}\sigma\bar{I}_L - \epsilon - \gamma - 2\mu\right)\hat{C}_{S_HI_L} \qquad \text{(B.54)}\end{aligned}$$

$$\begin{aligned}\frac{d\hat{C}_{S_LI_H}}{dt} &= -\frac{\beta}{N}\sigma\bar{S}_L C^H_{II} + \frac{\beta}{N}\sigma\bar{S}_H C^L_{SI} + \left(\frac{\beta}{N}(1-k\sigma)\bar{I}_H + k\frac{\beta}{N}\sigma\bar{I}_L + \epsilon\right)\hat{C}^H_{SS} - \frac{\beta}{N}(1-\sigma)\bar{S}_L\hat{C}^H_{II} \\ &\quad + \left(\frac{\beta}{N}(1-k\sigma)\bar{S}_H - \frac{\beta}{N}(1-\sigma)\bar{I}_L - \frac{\beta}{N}\sigma\bar{I}_H - \epsilon - \gamma - 2\mu\right)\hat{C}_{S_LI_H} + (k-1)\frac{\beta}{N}\sigma\bar{S}_H\hat{C}^L_{SI} \qquad \text{(B.55)}\end{aligned}$$

$$\begin{aligned}\frac{d\hat{C}^L_{SI}}{dt} &= \left(\frac{\beta}{N}(1-\sigma)\bar{I}_L + \frac{\beta}{N}\sigma\bar{I}_H + \epsilon\right)\hat{C}^L_{SS} - \frac{\beta}{N}\sigma\bar{S}_L\hat{C}^H_{II} - \frac{\beta}{N}(1-\sigma)\bar{S}_L\hat{C}^L_{II} + \frac{\beta}{N}\sigma\bar{S}_L\hat{C}_{S_LI_H} \\ &\quad + \left(\frac{\beta}{N}(1-\sigma)(\hat{S}_L - \hat{I}_L) - \frac{\beta}{N}\sigma\bar{I}_H - \epsilon - \gamma - 2\mu\right)\hat{C}^L_{SI}. \qquad \text{(B.56)}\end{aligned}$$

B.6 Derivation of the approximation for the star network

For the star network with P subpopulations on the star network, where the coupling between interacting populations is $\sigma \in [0, 1/k], k = P - 1$, we can show that the correlation between the number of infected individuals in the hub population and a leaf population, ρ_H, and the correlation between the number of infected individuals in two leaf populations, ρ_L are solution to the following pair of simultaneous equations:

$$\rho_H = \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}} \frac{\sigma}{\frac{\bar{S}_H^*}{\bar{S}_L^*}(\xi_H + k\sigma) + \xi_L + \sigma} + \sqrt{\frac{C_{II}^{L*}}{C_{II}^{H*}}} \frac{\sigma}{\xi_H + k\sigma + \frac{\bar{S}_L^*}{\bar{S}_H^*}(\xi_L + \sigma)} (1 - (1-k)\rho_L) - \Delta_H \tag{B.57}$$

$$\rho_L = \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}} \frac{\sigma}{\xi_L + \sigma} \rho_H + \Delta_L, \tag{B.58}$$

where

$$\xi_H = \frac{N(\gamma + \mu) - \beta \bar{S}_H^*}{\beta \bar{S}_H^*}, \tag{B.59}$$

$$\xi_L = \frac{N(\gamma + \mu) - \beta \bar{S}_L^*}{\beta \bar{S}_L^*} \tag{B.60}$$

and

$$\begin{aligned} \Delta_H = &\frac{\beta(1 - k\sigma)\bar{I}_H^* + k\beta\sigma\bar{I}_L^* + N\epsilon}{2N(\gamma + \mu) - \beta(1 - k\sigma)\bar{S}_H^* - \beta(1 - \sigma)\bar{S}_L^*} \frac{\hat{C}_{S_H I_L}^*}{\sqrt{C_{II}^{H*} C_{II}^{L*}}} \\ &+ \frac{\beta(1 - \sigma)\bar{I}_L^* + \beta\sigma\bar{I}_H^* + N\epsilon}{2N(\gamma + \mu) - \beta(1 - k\sigma)\bar{S}_H^* - \beta(1 - \sigma)\bar{S}_L^*} \frac{\hat{C}_{S_L I_H}^*}{\sqrt{C_{II}^{H*} C_{II}^{L*}}} \end{aligned} \tag{B.61}$$

$$\Delta_L = \frac{\beta(1 - \sigma)\bar{I}_L^* + \beta\sigma\bar{I}_H^* + N\epsilon}{N(\gamma + \mu) - \beta(1 - \sigma)\bar{S}_L^*} \frac{\hat{C}_{SI}^{L*}}{C_{II}^{L}}. \tag{B.62}$$

We derive Equation (B.57) using the moment equation for the covariance between the number of infected individuals in the hub and a leaf population, $\hat{C}_{II}^{H}$:

$$\begin{aligned}\frac{d\hat{C}_{II}^{H}}{dt} &= \frac{\beta}{N}\sigma\bar{S}_L C_{II}^{H} + \frac{\beta}{N}\sigma\bar{S}_H C_{II}^{L} + \left(\frac{\beta}{N}(1-k\sigma)\bar{S}_H + \frac{\beta}{N}(1-\sigma)\bar{S}_L - 2(\gamma+\mu)\right)\hat{C}_{II}^{H} \\ &\quad + (k-1)\frac{\beta}{N}\sigma\bar{S}_H\hat{C}_{II}^{L} + \left(\frac{\beta}{N}(1-k\sigma)\bar{I}_H + \frac{\beta}{N}k\sigma\bar{I}_L + \epsilon\right)\hat{C}_{S_H I_L} \\ &\quad + \left(\frac{\beta}{N}(1-\sigma)\bar{I}_L + \frac{\beta}{N}\sigma\bar{I}_H + \epsilon\right)\hat{C}_{S_L I_H}. \end{aligned} \tag{B.63}$$

At equilibrium $d\hat{C}_{II}^{H}/dt = 0$, and if we divide by $\sqrt{C_{II}^{H*}C_{II}^{L*}}/N$ then

$$\begin{aligned} 0 &= \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}}\beta\sigma\bar{S}_L^* + \sqrt{\frac{C_{II}^{L*}}{C_{II}^{H*}}}\beta\sigma\bar{S}_H^* + \left(\beta(1-k\sigma)\bar{S}_H^* + \beta(1-\sigma)\bar{S}_L^* - 2N(\gamma+\mu)\right)\rho_H \\ &\quad + \sqrt{\frac{C_{II}^{L*}}{C_{II}^{H*}}}(k-1)\beta\sigma\bar{S}_H^*\rho_L + \left(\beta(1-k\sigma)\bar{I}_H^* + \beta k\sigma\bar{I}_L^* + N\epsilon\right)\frac{\hat{C}_{S_H I_L}^*}{\sqrt{C_{II}^{H*}C_{II}^{L*}}} \\ &\quad + \left(\beta(1-\sigma)\bar{I}_L^* + \beta\sigma\bar{I}_H^* + N\epsilon\right)\frac{\hat{C}_{S_L I_H}^*}{\sqrt{C_{II}^{H*}C_{II}^{L*}}}, \end{aligned} \tag{B.64}$$

and hence we have the following approximation for the correlation between the number of infected individuals in the hub and a leaf population:

$$\begin{aligned}
\rho_H &= \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}}\frac{\beta\sigma\bar{S}_L^*}{2N(\gamma+\mu)-\beta(1-k\sigma)\bar{S}_H^*-\beta(1-\sigma)\bar{S}_L^*} \\
&\quad+\sqrt{\frac{C_{II}^{L*}}{C_{II}^{H*}}}\frac{\beta\sigma\bar{S}_H^*}{2N(\gamma+\mu)-\beta(1-k\sigma)\bar{S}_H^*-\beta(1-\sigma)\bar{S}_L^*}(1-(1-k)\rho_L)-\Delta_H \\
&= \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}}\frac{\sigma}{\frac{N(\gamma+\mu)-\beta(1-k\sigma)\bar{S}_H^*}{\beta\bar{S}_L^*}+\frac{N(\gamma+\mu)-\beta(1-\sigma)\bar{S}_L^*}{\beta\bar{S}_L^*}} \\
&\quad+\sqrt{\frac{C_{II}^{L*}}{C_{II}^{H*}}}\frac{\sigma}{\frac{N(\gamma+\mu)-\beta(1-k\sigma)\bar{S}_H^*}{\beta\bar{S}_H^*}+\frac{N(\gamma+\mu)-\beta(1-\sigma)\bar{S}_L^*}{\beta\bar{S}_H^*}}(1-(1-k)\rho_L)-\Delta_H \\
&= \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}}\frac{\sigma}{\frac{\bar{S}_H^*}{\bar{S}_L^*}\left(\frac{N(\gamma+\mu)-\beta\bar{S}_H^*}{\beta\bar{S}_H^*}+k\sigma\right)+\frac{N(\gamma+\mu)-\beta\bar{S}_L^*}{\beta\bar{S}_L^*}+\sigma} \\
&\quad+\sqrt{\frac{C_{II}^{L*}}{C_{II}^{H*}}}\frac{\sigma}{\frac{N(\gamma+\mu)-\beta\bar{S}_H^*}{\beta\bar{S}_H^*}+k\sigma+\frac{\bar{S}_L^*}{\bar{S}_H^*}\left(\frac{N(\gamma+\mu)-\beta\bar{S}_L^*}{\beta\bar{S}_L^*}+\sigma\right)}(1-(1-k)\rho_L)-\Delta_H \\
&= \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}}\frac{\sigma}{\frac{\bar{S}_H^*}{\bar{S}_L^*}(\xi_H+k\sigma)+\xi_L+\sigma}+\sqrt{\frac{C_{II}^{L*}}{C_{II}^{H*}}}\frac{\sigma}{\xi_H+k\sigma+\frac{\bar{S}_L^*}{\bar{S}_H^*}(\xi_L+\sigma)}(1-(1-k)\rho_L)-\Delta_H.
\end{aligned} \tag{B.65}$$

Similarly, we derive Equation (B.58) using the moment equations for the covariance between the number of infected individuals in two distinct leaf populations, $\hat{C}_{II}^L$:

$$\frac{d\hat{C}_{II}^L}{dt}=2\left[\frac{\beta}{N}\sigma\bar{S}_L\hat{C}_{II}^H+\left(\frac{\beta}{N}(1-\sigma)\bar{S}_L-\gamma-\mu\right)\hat{C}_{II}^L+\left(\frac{\beta}{N}(1-\sigma)\bar{I}_L+\frac{\beta}{N}\sigma\bar{I}_H+\epsilon\right)\hat{C}_{SI}^L\right]. \tag{B.66}$$

At equilibrium $d\hat{C}_{II}^L/dt=0$ and if we divide by $2C_{II}^{L*}/N$, then

$$0=\sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}}\beta\sigma\bar{S}_L^*\rho_H+\left(\beta(1-\sigma)\bar{S}_L^*-N(\gamma+\mu)\right)\rho_L+\left(\beta(1-\sigma)\bar{I}_L^*+\beta\sigma\bar{I}_H^*+N\epsilon\right)\frac{\hat{C}_{SI}^{L*}}{C_{II}^{L*}}, \tag{B.67}$$

and hence we have the following approximation for the correlation between the number of infected individuals in two leaf populations:

$$\begin{aligned}\rho_L &= \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}}\frac{\beta\sigma\bar{S}_L^*}{N(\gamma+\mu)-\beta(1-\sigma)\bar{S}_L^*}\rho_H + \frac{\beta(1-\sigma)\bar{I}_L^* + \beta\sigma\bar{I}_H^* + N\epsilon}{N(\gamma+\mu)-\beta(1-\sigma)\bar{S}_L^*}\frac{\hat{C}_{SI}^{L*}}{C_{II}^{L*}} \\ &= \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}}\frac{\sigma}{\frac{N(\gamma+\mu)-\beta\bar{S}_L^*}{\beta\bar{S}_L^*}+\sigma}\rho_H + \Delta_L \\ &= \sqrt{\frac{C_{II}^{H*}}{C_{II}^{L*}}}\frac{\sigma}{\xi_L+\sigma}\rho_H + \Delta_L. \end{aligned} \tag{B.68}$$

Bibliography

R. M. Anderson and R. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, UK, 1992.

H. Andersson and T. Britton. *Stochastic epidemic models and their statistical analysis*. Springer, 2000.

N. T. J. Bailey. *The Mathematical Theory of Infectious Diseases and Its Applications*. Charles Griffin \& Co Ltd, 1975.

D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.

D. Balcan, B. Gonçalves, H. Hu, J. J. Ramasco, V. Colizza, and A. Vespignani. Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. *Journal of Computational Science*, 1(3):132–145, 2010. ISSN 18777503. doi: 10.1016/j.jocs.2010.07.002.

F. Ball, T. Britton, T. House, V. Isham, D. Mollison, L. Pellis, and G. Scalia Tomba. Seven challenges for metapopulation models of epidemics, including households models. *Epidemics*, 10:63–67, 2014. ISSN 18780067. doi: 10.1016/j.epidem.2014.08.001.

A. D. Barbour. The principle of the diffusion of arbitrary constants. *Journal of Applied Probability*, 9 (3):519–541, 1972.

A. D. Barbour. On a functional central limit theorem for Markov population processes. *Advances in Applied Probability*, 6(1):21–39, 1974. ISSN 0001-8678. doi: 10.2307/1426205.

M. Barthélemy, C. Godreche, and J.-M. Luck. Fluctuation effects in metapopulation models: Percolation and pandemic threshold. *Journal of Theoretical Biology*, pages 554–564, 2010. doi: 10.1016/j.jtbi.2010.09.015.

V. Belik, T. Geisel, and D. Brockmann. Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X*, 1:1–5, 2011. doi: 10.1103/PhysRevX.1.011001.

M. Biggerstaff, S. Cauchemez, C. Reed, M. Gambhir, and L. Finelli. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC infectious Diseases*, 14:1–20, 2014.

B. M. Bolker. Chaos and complexity in measles models: a comparative numerical study. *IMA Journal of Mathematics Applied in Medicine and Biology*, 10(2):83–95, 1993. ISSN 14778599. doi: 10.1093/imammb/10.2.83.

B. M. Bolker and B. T. Grenfell. Impact of vaccination on the spatial correlation and persistence of measles dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 93(22):12648–12653, 1996. ISSN 0027-8424. doi: 10.1073/pnas.93.22.12648.

Y. Cao, D. T. Gillespie, and L. R. Petzold. Efficient step size selection for the tau-leaping simulation method. *Journal of Chemical Physics*, 124(4), 2006. ISSN 00219606. doi: 10.1063/1.2159468.

K. Capała and B. Dybiec. Epidemics spread in heterogeneous populations. *European Physical Journal B*, 90(5), 2017. ISSN 14346036. doi: 10.1140/epjb/e2017-70723-6.

C. Cattuto, W. van den Broeck, A. Barrat, V. Colizza, J. F. Pinton, and A. Vespignani. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE*, 5(7), 2010. ISSN 19326203. doi: 10.1371/journal.pone.0011596.

S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron, and P. Y. Boelle. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine*, 23:3469–3487, 2004. doi: 10.1002/sim.1912.

S. Cauchemez, A. Bhattarai, T. L. Marchbanks, R. P. Fagan, S. Ostroff, N. M. Ferguson, D. Swerdlow, S. V. Sodha, M. E. Moll, F. J. Angulo, R. Palekar, W. R. Archer, and L. Finelli. Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7): 2825–2830, 2011. ISSN 00278424. doi: 10.1073/pnas.1008895108.

R. M. Christley, G. L. Pinchbeck, R. G. Bowers, D. Clancy, N. P. French, and R. Bennett. Infection in social networks: using network analysis to identify high-risk individuals. *American Journal of Epidemiology*, 162(10):1024–1031, 2005. doi: 10.1093/aje/kwi308.

V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences*, 103(7):2015–2020, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0510525103.

L. Danon, A. P. Ford, T. House, C. P. Jewell, M. J. Keeling, G. O. Roberts, J. V. Ross, and M. C. Vernon. Networks and the epidemiology of infectious disease. *Interdisciplinary Perspectives on Infectious Diseases*, 2011:28, 2011. ISSN 1687708X. doi: 10.1155/2011/284909.

L. Danon, J. M. Read, T. A. House, M. C. Vernon, and M. J. Keeling. Social encounter networks: characterizing Great Britain. *Proceedings of the Royal Society B*, 2013.

G. De Serres, F. Markowski, E. Toth, M. Landry, D. Auger, M. Mercier, P. Bélanger, B. Turmel, H. Arruda, N. Boulianne, B. J. Ward, and D. M. Skowronski. Largest measles epidemic in North America in a decade-Quebec, Canada, 2011: Contribution of susceptibility, serendipity, and superspreading events. *Journal of Infectious Diseases*, 207(6):990–998, 2013. ISSN 00221899. doi: 10.1093/infdis/jis923.

O. Diekmann and J. A. P. Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. John Wiley \& Sons, 2000.

O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28:365–382, 1990.

L. Dyson, W. A. Stolk, S. H. Farrell, and T. D. Hollingsworth. Measuring and modelling the effects of systematic non-adherence to mass drug administration. *Epidemics*, 18:56–66, 2017. ISSN 18780067. doi: 10.1016/j.epidem.2017.02.002.

D. J. D. Earn, P. Rohani, B. T. Grenfell, and B. M. Bolker. A simple model for complex dynamical transitions in epidemics. *Science*, 287(5453):667–670, 2000.

S. Erlander and N. F. Stewart. *The Gravity Model in Transportation Analysis – Theory and Extensions*. 1990.

D. T. Gillespie. A general method for numerically simulation the stochastic time evolution of coupled chemical reactions, 1976. ISSN 10902716.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. In *Journal of Physical Chemistry*, volume 81, pages 2340–2361, 1977. doi: 10.1021/j100540a008.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115(4):1716–1733, 2001. ISSN 00219606. doi: 10.1063/1.1378322.

D. T. Gillespie. Stochastic Simulation of Chemical Kinetics. *Annual Review of Physical Chemistry*, 58 (1):35–55, 2007. ISSN 0066-426X. doi: 10.1146/annurev.physchem.58.032806.104637.

D. T. Gillespie and L. R. Petzold. Improved lead-size selection for accelerated stochastic simulation. *Journal of Chemical Physics*, 119(16):8229–8234, 2003. ISSN 00219606. doi: 10.1063/1.1613254.

J. R. Gog, S. Ballesteros, C. Viboud, L. Simonsen, O. N. Bjornstad, J. Shaman, D. L. Chao, F. Khan, and B. T. Grenfell. Spatial Transmission of 2009 Pandemic Influenza in the US. *PLoS Computational Biology*, 10(6):1003635, 2014. ISSN 15537358. doi: 10.1371/journal.pcbi.1003635.

M. C. González, C. A. Hidalgo, and A. L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008. ISSN 14764687. doi: 10.1038/nature06958.

B. T. Grenfell and R. M. Anderson. The estimation of age related rates of infection from case notifications and serological data. *Epidemiology \& Infection*, 95(2):419–436, 1985.

B. T. Grenfell and B. M. Bolker. Spatial heterogeneity, nonlinear dynamics and chaos in infectious diseases. *Statistical Methods in Medical Research*, 4(2):160–183, 1995.

B. T. Grenfell and J. Harwood. (Meta)population dynamics of infectious diseases. *Trends in Ecology & Evolution*, 12(10):395–399, 1997. ISSN 01695347. doi: 10.1016/S0169-5347(97)01174-9.

G. Grimmet and D. Stirzaker. *Probability and Random Processes*. 2001.

R. Guimerà, S. Mossa, A. Turtschi, and L. A. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(22):7794–7799, 2005. ISSN 00278424. doi: 10.1073/pnas.0407994102.

T. J. Hagenaars, C. A. Donnelly, and N. M. Ferguson. Spatial heterogeneity and the persistence of infectious diseases. *Journal of Theoretical Biology*, 229:349–359, 2004. doi: 10.1016/j.jtbi.2004.04.002.

I. Hanski and M. Gilpin. Metapopulation dynamics: brief history and conceptual domain. *Biological Journal of the Linnean Society*, 1991.

I. Hanski and D. Simberloff. The Metapopulation Approach, Its History, Conceptual Domain, and Application to Conservation. In *Metapopulation Biology*, pages 5–26. 1997. doi: 10.1016/b978-012323445-2/50003-1.

I. A. Hanski. Metapopulation dynamics. *Nature*, 396:41–49, 1998.

J. A. P. Heesterbeek. A brief history of R0 and a recipe for its calculation. *Acta Biotheoretica*, 50(3): 189–204, 2002. ISSN 00015342. doi: 10.1023/A:1016599411804.

J. Hilton and M. J. Keeling. Incorporating household structure and demography into models of endemic disease. *Journal of The Royal Society Interface*, 2019. doi: 10.1098/rsif.2019.0317.

P. Horby, P. Q. Thai, N. Hens, N. T. T. Yen, L. Q. Mai, D. D. Thoang, N. M. Linh, N. T. Huong, N. Alexander, W. J. Edmunds, T. N. Duong, A. Fox, and N. T. Hien. Social contact patterns in Vietnam and implications for the control of infectious diseases. *PLoS ONE*, 6(2), 2011. ISSN 19326203. doi: 10.1371/journal.pone.0016965.

V. Isham. Assessing the variability of stochastic epidemics. *Mathematical Biosciences*, 107(2):209–224, 1991. ISSN 00255564. doi: 10.1016/0025-5564(91)90005-4.

V. Isham. Stochastic models for epidemics with special reference to AIDS. *The Annals of Applied Probability*, 3(1):1–27, 1993.

V. Isham. Stochastic models of host-macroparasite interaction. *The Annals of Applied Probability*, 5(3): 720–740, 1995.

J. A. Jacquez and C. P. Simon. The stochastic SI model with recruitment and deaths I. comparison with the closed SIS model. *Mathematical Biosciences*, 117(1-2):77–125, 1993. ISSN 00255564. doi: 10.1016/0025-5564(93)90018-6.

M. Jesse, P. Ezanno, S. Davis, and J. A. Heesterbeek. A fully coupled, mechanistic model for infectious disease dynamics in a metapopulation: Movement and epidemic duration. *Journal of Theoretical Biology*, 254(2):331–338, 2008. ISSN 00225193. doi: 10.1016/j.jtbi.2008.05.038.

C. Kang, Y. Liu, D. Guo, and K. Qin. A generalized radiation model for human mobility: Spatial scale, searching direction and trip constraint. *PLoS ONE*, 10(11), 2015. ISSN 19326203. doi: 10.1371/journal.pone.0143500.

M. J. Keeling. Metapopulation moments: Coupling, stochasticity and persistence. *Journal of Animal Ecology*, 69(5):725–736, 2000a. ISSN 00218790. doi: 10.1046/j.1365-2656.2000.00430.x.

M. J. Keeling. Multiplicative moments and measures of persistence in ecology. *Journal of Theoretical Biology*, 205(2):269–281, 2000b. ISSN 0022-5193. doi: 10.1006/jtbi.2000.2066.

M. J. Keeling and B. T. Grenfell. Disease extinction and community size: modeling the persistence of measles. *Science*, 275(5296):65–67, 1997.

M. J. Keeling and P. Rohani. Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecology Letters*, 5(1):20–29, 2002. ISSN 1461023X. doi: 10.1046/j.1461-0248.2002.00268.x.

M. J. Keeling and P. Rohani. *Modelling Infectious Diseases in Humans and Animals*. Princeton University Press, 2008.

M. J. Keeling and J. V. Ross. On methods for studying stochastic disease dynamics. *Journal of the Royal Society Interface*, 5(19):171–181, 2008. ISSN 17425662. doi: 10.1098/rsif.2007.1106.

M. J. Keeling and P. J. White. Targeting vaccination against novel infections: risk, age and spatial structure for pandemic influenza in Great Britain. *Journal of The Royal Society Interface*, 8(58): 661–670, 2010. ISSN 1742-5689. doi: 10.1098/rsif.2010.0474.

W. O. Kermack and A. G. McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 115(772):700–721, 1927. ISSN 10991476. doi: 10.1002/mma.5067.

M. C. Kiti, M. Tizzoni, T. M. Kinyanjui, D. C. Koech, P. K. Munywoki, M. Meriac, L. Cappa, A. Panisson, A. Barrat, C. Cattuto, and D. J. Nokes. Quantifying social contacts in a household setting of rural Kenya using wearable proximity sensors. *EPJ Data Science*, 5(1), 2016. ISSN 21931127. doi: 10.1140/epjds/s13688-016-0084-2.

P. Klepac, S. Kissler, and J. Gog. Contagion! The BBC Four Pandemic The model behind the documentary. *Epidemics*, 24:49–59, 2018. ISSN 18780067. doi: 10.1016/j.epidem.2018.03.003.

A. Kolmogoroff. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Mathematische Annalen*, 104(1):415–458, 1931. ISSN 00255831. doi: 10.1007/BF01457949.

M. U. Kraemer, N. R. Faria, R. C. Reiner, N. Golding, B. Nikolay, S. Stasse, M. A. Johansson, H. Salje, O. Faye, G. R. Wint, M. Niedrig, F. M. Shearer, S. C. Hill, R. N. Thompson, D. Bisanzio, N. Taveira, H. H. Nax, B. S. Pradelski, E. O. Nsoesie, N. R. Murphy, I. I. Bogoch, K. Khan, J. S. Brownstein, A. J. Tatem, T. de Oliveira, D. L. Smith, A. A. Sall, O. G. Pybus, S. I. Hay, and S. Cauchemez. Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 201516: a modelling study. *The Lancet Infectious Diseases*, 17(3):330–338, 2017. ISSN 14744457. doi: 10.1016/S1473-3099(16)30513-8.

I. Krishnarajah, A. Cook, G. Marion, and G. Gibson. Novel moment closure approximations in stochastic epidemics. *Bulletin of Mathematical Biology*, 67:855–873, 2005. doi: 10.1016/j.bulm.2004.11.002.

A. J. Kucharski and C. L. Althaus. The role of superspreading in middle east respiratory syndrome coronavirus (Mers-CoV) transmission. *Eurosurveillance*, 20(25):1–5, 2015. ISSN 15607917.

T. G. Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability*, 7(1):49–58, 1970.

T. G. Kurtz. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 8(2):344–356, 1971.

G. E. Lahodny and L. J. Allen. Probability of a Disease Outbreak in Stochastic Multipatch Epidemic Models. *Bulletin of Mathematical Biology*, 75(7):1157–1180, 2013. ISSN 00928240. doi: 10.1007/s11538-013-9848-z.

M. S. Lau, B. D. Dalziel, S. Funk, A. McClelland, A. Tiffany, S. Riley, C. J. E. Metcalf, and B. T. Grenfell. Spatial and temporal dynamics of superspreading events in the 2014-2015 West Africa Ebola epidemic. *Proceedings of the National Academy of Sciences of the United States of America*, 114(9):2337–2342, 2017. ISSN 10916490. doi: 10.1073/pnas.1614595114.

R. Levins. Some Demographic and Genetic Consequences of Environmental Heterogeneity for Biological Control. *Bulletin of the Entomological Society of America*, 15(3):237–240, 1969. ISSN 0013-8754. doi: 10.1093/besa/15.3.237.

A. L. Lloyd. Estimating variability in models for recurrent epidemics: assessing the use of moment closure techniques. *Theoretical Population Biology*, 65(1):49–65, 2004. ISSN 00405809. doi: 10.1016/j.tpb.2003.07.002.

A. L. Lloyd and V. A. Jansen. Spatiotemporal dynamics of epidemics: Synchrony in metapopulation models. In *Mathematical Biosciences*, volume 188, pages 1–16, 2004. doi: 10.1016/j.mbs.2003.09.003.

A. L. Lloyd and R. M. May. Spatial heterogeneity in epidemic models. *Journal of theoretical biology*, 179(1):1–11, 1996. ISSN 0022-5193. doi: 10.1006/jtbi.1996.0042.

J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359, 2005. ISSN 14764687. doi: 10.1038/nature04153.

J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, and W. J. Edmunds. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, 5(3):381–391, 2008. ISSN 15491277. doi: 10.1371/journal.pmed.0050074.

I. Nasell. On the time to extinction in recurrent epidemics. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(2):309–330, 1999.

I. Nasell. Moment closure and the stochastic logistic model. *Theoretical Population Biology*, 63(2): 159–168, 2003a. ISSN 00405809. doi: 10.1016/S0040-5809(02)00060-6.

I. Nasell. An extension of the moment closure method. *Theoretical Population Biology*, 64:233–239, 2003b. doi: 10.1016/S0040-5809(03)00074-1.

L. F. Olsen and W. M. Schaffer. Chaos versus noisy periodicity: alternative hypotheses for childhood epidemics. *Science*, 249(4968):499–504, 1990. doi: 10.1126/science.2382131.

M. Rathinam, L. R. Petzold, Y. Cao, and D. T. Gillespie. Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. *Journal of Chemical Physics*, 119(24):12784–12794, 2003. ISSN 00219606. doi: 10.1063/1.1627296.

H. F. Raymond and W. McFarland. Racial mixing and HIV risk among men who have sex with men. *AIDS and Behavior*, 13(4):630–637, 2009. ISSN 10907165. doi: 10.1007/s10461-009-9574-6.

J. M. Read, J. Lessler, S. Riley, S. Wang, L. J. Tan, K. O. Kwok, Y. Guan, C. Q. Jiang, and D. A. T. Cummings. Social mixing patterns in rural and urban areas of southern China. *Proceedings of the Royal Society B*, 281(1785), jun 2014.

P. Rodrigues, A. Margheri, C. Rebelo, and M. G. M. Gomes. Heterogeneity in susceptibility to infection can explain high reinfection rates. *Journal of Theoretical Biology*, 259(2):280–290, 2009. ISSN 10958541. doi: 10.1016/j.jtbi.2009.03.013.

P. Rohani, M. J. Keeling, and B. T. Grenfell. The interplay between determinism and stochasticity in childhood diseases. *The American Naturalist*, 159(5):469–481, 2002. ISSN 0003-0147. doi: 10.1086/339467.

G. Rozhnova, A. Nunes, and A. J. Mckane. Phase lag in epidemics on a network of cities. *Physical Review E*, 85(5):051912, 2012. doi: 10.1103/PhysRevE.85.051912.

L. Sattenspiel and K. Dietz. A structured epidemic model incorporating geographic mobility among regions. *Mathematical Biosciences*, 128(1):71–91, 1995.

D. Schenzle. An age-structured model of pre- and post-vaccination measles transmission. *Mathematical Medicine and Biology: A Journal of the IMA*, 1(2):169–191, 1984.

J. A. Schneider, B. Cornwell, D. Ostrow, S. Michaels, P. Schumm, E. O. Laumann, and S. Friedman. Network mixing and network influences most linked to HIV infection and risk behavior in the HIV epidemic among black men who have sex with men. *American Journal of Public Health*, 103(1):28–36, 2013. ISSN 00900036. doi: 10.2105/AJPH.2012.301003.

J. A. G. Scott, E. Bauni, J. C. Moisi, J. Ojal, H. Gatakaa, C. Nyundo, C. S. Molyneux, F. Kombe, B. Tsofa, K. Marsh, N. Peshu, and T. N. Williams. Profile: The Kilifi health and demographic surveillance system (KHDSS). *International Journal of Epidemiology*, 41(3):650–657, 2012. ISSN 03005771. doi: 10.1093/ije/dys062.

T. Sellke. On the Asymptotic Distribution of the Size of a Stochastic Epidemic. *Journal of Applied Probability*, 20(2):390–394, 1983. ISSN 00219002. doi: 10.2307/3213811.

F. Simini, M. C. González, A. Maritan, and A. L. Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012. ISSN 00280836. doi: 10.1038/nature10856.

J. Stehlé, F. Charbonnier, T. Picard, C. Cattuto, and A. Barrat. Gender homophily from spatial behavior in a primary school: A sociometric study. *Social Networks*, 35(4):604–613, 2013. ISSN 03788733. doi: 10.1016/j.socnet.2013.08.003.

A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, J. E. Larsen, and S. Lehmann. Measuring large-scale social networks with high resolution. *PLoS ONE*, 9(4):95978, 2014. ISSN 19326203. doi: 10.1371/journal.pone.0095978.

M. Tizzoni, P. Bajardi, A. Decuyper, G. Kon, K. King, and C. M. Schneider. On the use of human mobility proxies for modeling epidemics. *PLoS Computational Biology*, 10(7), 2014. doi: 10.1371/journal.pcbi.1003716.

C. Viboud, O. N. Bjørnstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*, 312(5772):447–51, 2006. ISSN 1095-9203. doi: 10.1126/science.1125237.

J. Wallinga, M. van Boven, and M. Lipsitch. Optimizing infectious disease interventions during an emerging epidemic. *Proceedings of the National Academy of Sciences*, 107(2):923–928, jan 2010.

L. Wang and J. T. Wu. Characterizing the dynamics underlying global spread of epidemics. *Nature Communications*, 9(1), 2018. ISSN 20411723. doi: 10.1038/s41467-017-02344-z.

A. Wesolowski, W. P. O'Meara, N. Eagle, A. J. Tatem, and C. O. Buckee. Evaluating Spatial Interaction Models for Regional Mobility in Sub-Saharan Africa. *PLoS Computational Biology*, 11(7):1004267, 2015. ISSN 15537358. doi: 10.1371/journal.pcbi.1004267.

P. Whittle. On the use of the normal approximation in the treatment of stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 19(2):268–281, 1957.

Z. Wu, K. Rou, and H. Cui. The HIV/AIDS epidemic in China: History, current strategies and future challenges. *AIDS Education and Prevention*, 16(3 SUPPL.):7–17, 2004. ISSN 08999546.

A. W. Yan, A. J. Black, J. M. McCaw, N. Rebuli, J. V. Ross, A. J. Swan, and R. I. Hickson. The distribution of the time taken for an epidemic to spread between two communities. *Mathematical Biosciences*, 303:139–147, jul 2018. ISSN 18793134. doi: 10.1016/j.mbs.2018.07.004.

X. Y. Yan, C. Zhao, Y. Fan, Z. Di, and W. X. Wang. Universal predictability of mobility patterns in cities. *Journal of the Royal Society Interface*, 11(100), 2014. ISSN 17425662. doi: 10.1098/rsif.2014.0834.